Variational Locally Projected Regression

Master thesis by Tom Friedrich Buchholz
Date: December 6, 2021, Date of submission: December 6, 2021

1. Review: Hany Abdulsamad
2. Review: Janosch Moos
3. Review: Deborah Clever
4. Review: Jan Peters

Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Computer Science Department

Intelligent Autonomous Systems

Variational Locally Projected Regression

Master thesis by Tom Friedrich Buchholz

1. Review: Hany Abdulsamad
2. Review: Janosch Moos
3. Review: Deborah Clever
4. Review: Jan Peters

Date: December 6, 2021
Date of submission: December 6, 2021

Darmstadt

Bitte zitieren Sie dieses Dokument als:
URN: urn:nbn:de:tuda-tuprints-1234
URL: http://tuprints.ulb.tu-darmstadt.de/12345
DOI: https://doi.org/10.25534/tuprints-1234

Dieses Dokument wird bereitgestellt von tuprints,
E-Publishing-Service der TU Darmstadt
http://tuprints.ulb.tu-darmstadt.de
tuprints@ulb.tu-darmstadt.de

Erklärung zur Abschlussarbeit gemäß §22 Abs. 7 und §23 Abs. 7 APB der TU Darmstadt

Hiermit versichere ich, Tom Friedrich Buchholz, die vorliegende Masterarbeit ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Fall eines Plagiats (§38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Bei der abgegebenen Thesis stimmen die schriftliche und die zur Archivierung eingereichte elektronische Fassung gemäSS §23 Abs. 7 APB überein.

Bei einer Thesis des Fachbereichs Architektur entspricht die eingereichte elektronische Fassung dem vorgestellten Modell und den vorgelegten Plänen.

Darmstadt, December 6, 2021

Tom Friedrich Buchholz

Abstract

In this thesis we develop a variational locally projected regression model. This work is based on the ideas of principal component analysis, and its probabilistic formulation and extension to mixture models. The general idea of this thesis is that many datasets have an underlying structure that is difficult to detect in the data space itself. Doing regression on such datasets requires overly complex models with redundant parameters and leads to suboptimal approximations. In this approach, the dataset is transformed into a higher dimensional latent space using nonlinear projections, such that the representation of the data and its underlying structure simplifies. A mixture of probabilistic linear models is trained on the projected data with variational inference, resulting in a bayesian approximation with uncertainty quantification and reduced parameter redundancy.

Zusammenfassung

In dieser Arbeit entwickeln wir ein variationales lokal projiziertes Regressionsmodell. Diese Arbeit basiert auf den Ideen der Hauptkomponentenanalyse und ihrer probabilistischen Formulierung und Erweiterung auf Mischmodelle. Der Grundgedanke dieser Arbeit ist, dass viele Datensätze eine zugrundeliegende Struktur haben, die im Datenraum selbst schwer zu erkennen ist. Die Durchführung von Regressionen auf solchen Datensätzen erfordert übermäßig komplexe Modelle mit redundanten Parametern und führt zu suboptimalen Näherungen. Bei diesem Ansatz wird der Datensatz mit Hilfe nichtlinearer Projektionen in einen höherdimensionalen latenten Raum transformiert, so dass sich die Darstellung der Daten und ihrer zugrunde liegenden Struktur vereinfacht. Eine Mischung probabilistischer linearer Modelle wird auf den projizierten Daten mit Variationsinferenz trainiert, was zu einer bayesianischen Annäherung mit Quantifizierung der Unsicherheit und reduzierter Parameterredundanz führt.

Contents

1 Introduction and Overview

With the development of affordable computers and rising processing power, artificial intelligence and machine learning made considerable advancements in solving previously intractable problems. The resulting technologies are increasingly important in our lives and are adapted in various fields.

With the increasing number and size of available datasets, deep neural networks became one of those advancements. They are easy-to-train universal approximators that can handle high-dimensional and complex problems while being performant and achieving exceptional results. A significant disadvantage of neural networks is that it is not well understood how the learned parameters are related to the given task. They are also prone to over-parametrization [1], [9], catastrophic forgetting [15], and unseen data can lead to unexpected behavior [11].

In contrast, probabilistic methods like gaussian process regression (GPR) [21] offer a principled Bayesian treatment that allows to counteract against over-parametrization with Automatic Relevance Detection (ARD) [10] and offer uncertainty quantification [17] for their results. There are, however, several downsides to methods that are using the Bayesian framework. They are generally computationally more expensive and often have many hyperparameters, making them difficult to tune and limiting their application to complex tasks.

Another approach has been to simplify the dataset by combining multiple simpler models such as mixture models [8] , as well as discarding dimensions by projecting the data onto a lower-dimensional subspace [20], [5], [3]. Splitting the regression problem into several sub-problems simplifies the training even with a Bayesian approach. However, a downside is that data often contains repeating structures that may not be captured when using several local approximators, leading to redundant parameters and inefficient learning.

In this thesis, we try to make use of similar structures in the dataset by projecting the datapoints onto a higher-dimensional latent space, such that the structure simplifies under the projection. We then utilize a mixture of linear probabilistic models and use Variational Inference (VI) to do regression on the projected data. This approach seeks to combine the benefits of the Bayesian framework while being able to approximate complex high-dimensional data and keeping the computational cost and parameter redundancy low.

In the following chapters, several foundational techniques are introduced, starting with Principal Component Analysis (PCA) in Section 2.1.

PCA is a widely used globally linear method that reduces the dimensionality of a given dataset. In order to diminish its linear restriction, we then introduce the probabilistic extension of PCA in Section 2.2, which enables a principled way of incorporating mixture models.

To eliminate the need to specify the dimensionality of the projected data, we show how to augment Probablistic Principal Component Analysis (PPCA) to a bayesian formulation in Section 2.3. With Bayesian Principal Component Analysis (BPCA) we are able to use ARD to infer the dimensionality-hyperparameter from the data.

In Section 2.4 we give an overview of the Expectation Maximization (EM) algorithm used to efficiently learn the parameters of probabilistic models.

We then address variations of PPCA and BPCA that work with mixture models in Chapter 3. The variations Mixture of Probabilistic Principal Component Analysers (MPPCA) and Bayesian Mixture Principal Component Analysis (BMPCA) can be applied to more complex problems and build the foundations for the method proposed in the next chapter.

Chapter 4 motivates the idea of this thesis, proposes a probabilistic model with a hidden layer and demonstrates how to derive the learning algorithm with VI.

We end with a summary and propositions for future work.

2 Foundations

In this chapter several foundational technologies are introduced, that form the basic building blocks to understand the proposed new method.

2.1 Principal Component Analysis

PCA is a technique capable of reducing the dimensionality of data, while minimizing the loss of information [12,13,20]. It uses eigendecomposition to find an orthogonal coordinate system, that is aligned with the directions of maximal variance in the data. The data is then projected onto these principal axes, while omitting dimensions with the least amount of variance. This results in an approximate representation of the data, with a lower dimensionality. The linear subspace is called principal subspace. While PCA is a globally linear method, we will later lift that restriction.

Reducing the dimensionality of a given dataset in a machine learning task, prevents learning characteristics in the dataset, that have little influence on the given task, helps reducing the complexity of the learned model and makes the learning process more feasible.

Let D be the dimension of the given data, and Q the dimension of the principal subspace, where $Q < D$. And furthermore we define $\{\mathbf{t}_n\}, n \in \{1..N\}$ as the D-dimensional observations, $\{\mathbf{x}_n\}, n \in \{1..N\}$ as the Q-dimensional projected data, and $\mathbf{w}_j, j \in \{1..q\}$ as vectors representing the principal axes.

Given N datapoints $\mathbf{t}_n$, we first calculate the sample mean by

$$\bar{\mathbf{t}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{t}_n.$$

Using the sample mean, we can define the sample covariance matrix $\mathbf{S}$, which correlates the variance of the data with the axes of the coordinate system of the dataset, with

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{t}_n - \bar{\mathbf{t}})(\mathbf{t}_n - \bar{\mathbf{t}})^\top.$$

By using eigendecomposition, we can derive the eigenvectors $\mathbf{w}_i$ and eigenvalues $\boldsymbol{\lambda}_i$ of the sample covariance matrix $\mathbf{S}$, with

$$\mathbf{S}\mathbf{w}_i = \boldsymbol{\lambda}_i \mathbf{w}_i.$$

The eigenvalue λ_i describes the variance of the data, along the corresponding eigenvector $\mathbf{w}_i$.

Now we omit $D-Q$ dimensions by choosing the Q largest (where $Q < D$) eigenvalues and their corresponding eigenvectors and define a projection matrix $\mathbf{W}$, where each column is an eigenvector $\mathbf{w}_i$, written as

$$\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_q).$$

The projection matrix $\mathbf{W}$ can be used to project the data on the principal subspace, by minimizing the sum-of-squares reconstruction cost. The projected $\mathbf{x}_n$ are defined as

$$\mathbf{x}_n = \mathbf{W}^T(\mathbf{t}_n - \bar{\mathbf{t}}).$$

The lower dimensional approximate data representation $\mathbf{x}$ can be used to reconstruct the data in the original coordinate space using the inverse projection [6]. The reconstructed datapoints $\hat{\mathbf{t}}_n$ are given by

$$\hat{\mathbf{t}}_n = \mathbf{W}\mathbf{x}_n + \bar{\mathbf{t}}.$$

In Figure 2.1 on page 5, we see observations (in blue), that are sampled from a skewed two-dimensional Gaussian distribution, and the reconstructed datapoints $\hat{\mathbf{t}}_n$ in yellow. We can see that on the omitted dimension, information is lost.

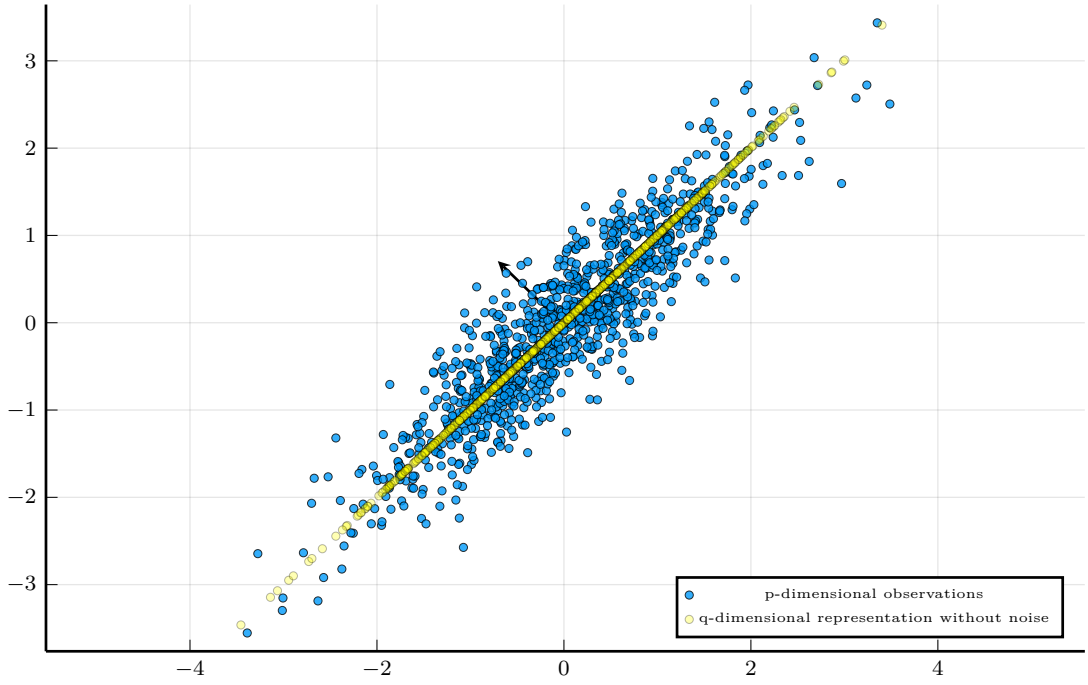

Figure 2.1: PCA applied to a dataset, sampled from a skewed two-dimensional Gaussian distribution. The samples, shown in blue, are projected onto a one-dimensional principal subspace, which principal axis rotated in the direction with the biggest variance. The projected datapoints are then projected back into the original space, and being shown in yellow.

2.2 Probabilistic Principal Component Analysis

PPCA is a probabilistic formulation of PCA as proposed by [18] and [20] , giving several advantages over the previously introduced non probabilistic version of PCA.

With a probabilistic model, we gain the ability to generate datapoints, deal with missing data and apply it to classification problems. We can also find solutions for high dimensional data more efficiently, using the EM algorithm. Furthermore we can combine multiple sub models into a mixture model in a principled way [6]. If the covariance $\sigma^2\mathbf{I}$ would be defined as a diagonal matrix $\mathbf{\Psi}$ where the diagonal values can differ, this algorithm is called factor analysis [16].

We again look at a dataset of dimension D, consisting of N datapoints $\mathbf{t}_n$, and introduce a Q-dimensional latent variable $\mathbf{x}_n$ for every $\mathbf{t}_n$.

The dimensionality of the latent space is again supposed to be lower than the data space: $Q < D$.

A general transformation between the latent- and dataspace with added noise is given by

$$\mathbf{t} = \mathbf{y}(\mathbf{x}; \mathbf{w}) + \boldsymbol{\epsilon}.$$

In this case we use a linear transformation, with the projection matrix $\mathbf{W}$, an offset $\boldsymbol{\mu}$ and added noise $\boldsymbol{\epsilon}$, we can formulate

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{2.1}$$

where the projection matrix $\mathbf{W}$ is a $D \times Q$ matrix, whose column vectors span the principal subspace.

For a probabilistic treatment, we need to introduce a prior distribution over the latent variable $\mathbf{x}$ with

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}).$$

The conditional distribution is also gaussian, with a mean given by the linear transformation (2.1), and for the variance we introduce a new parameter σ, resulting in

$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}_D).$$

Given the prior- and conditional distribution, we can integrate over the latent variable $\mathbf{x}$ to obtain the marginal distribution over the observed data $\mathbf{t}$. Since we use a conjugate prior distribution, the marginal is a gaussian with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$, given by

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\, d\mathbf{x} = \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \mathbf{C}). \tag{2.2}$$

Under the assumption that $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are independent, we can derive the mean of (2.2) by using the expectation [20]

$$\mathbb{E}[\mathbf{t}] = \mathbb{E}[\mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}.$$

We see that the mean of the marginal distribution is given by the offset parameter $\boldsymbol{\mu}$ in (2.1).

For the covariance $\mathbf{C}$ we find

$$\begin{aligned}\mathbf{C} &= \mathbb{E}[(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})^\top] \\ &= \mathbb{E}[\mathbf{W}\mathbf{x}\mathbf{x}^\top\mathbf{W}^\top] + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] \\ &= \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}_D.\end{aligned}$$

Instead of calculating the inverse of the covariance matrix $\mathbf{S}$ with dimension $D \times D$ directly, we use the matrix inversion identity [6]

$$\mathbf{C}^{-1} = \sigma^2\mathbf{I} - \sigma^{-2}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T,$$

where

$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}. \tag{2.3}$$

This transformation simplifies the inversion of the $D \times D$ matrix $\mathbf{S}$, to an inversion of the $Q \times Q$ matrix $\mathbf{M}$, which can be a significant computational advantage on high-dimensional datasets.

The predictive distribution can now be written as

$$p(\mathbf{x}|\mathbf{t}) = \mathcal{N}(\mathbf{x}|\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{t} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}).$$

In the next section, the maximum likelihood solutions are shown.

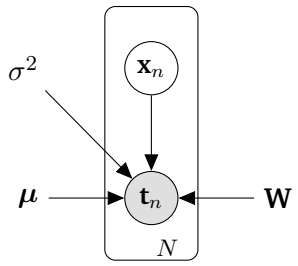

Figure 2.2: Graphical model of PPCA. Each observation $\mathbf{t}_n$ is modeled by a directed graph with the mean $\boldsymbol{\mu}$, variance σ^2 and projection $\mathbf{W}$. For each $\mathbf{t}_n$ there is a corresponding $\mathbf{x}_n$.

2.2.1 Maximum Likelihood Solutions

In order to approximate the given data, with the derived model, we need to find the optimal parameters $\boldsymbol{\mu}$, σ and $\mathbf{W}$.

Defining the log likelihood function as

$$\begin{aligned}
\ln p(\mathbf{T}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) &= \sum_{n=1}^{N} \ln p(\mathbf{t}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) \\
&= \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}, \mathbf{C}) \\
&= \sum_{n=1}^{N} \ln \left[(2\pi)^{-D/2} |\mathbf{C}|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{t}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1} (\mathbf{t}_n - \boldsymbol{\mu}) \right\} \right] \\
&= \sum_{n=1}^{N} \left[-\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{C}| - \frac{1}{2}(\mathbf{t}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1} (\mathbf{t}_n - \boldsymbol{\mu}) \right] \\
&= -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\mathbf{C}| - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{t}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1} (\mathbf{t}_n - \boldsymbol{\mu}),
\end{aligned}$$

we can set the derivatives w.r.t the parameters equal to 0, and solve the equation for the corresponding parameter, in order to find the maximum likelihood solutions.

The mean $\boldsymbol{\mu}$ is defined as the mean of the dataset

$$\boldsymbol{\mu}_{ML} = \bar{\mathbf{t}}.$$

Substituting the result for $\boldsymbol{\mu}$ back into the log likelihood function and using the sample covariance matrix $\mathbf{S}$

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \bar{\mathbf{t}})(\mathbf{t}_n - \bar{\mathbf{t}})^\top,$$

the log likelihood simplifies to

$$\begin{aligned}\ln p(\mathbf{T}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) &= -\frac{N}{2}\big\{d\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}[\mathbf{C}^{-1}\mathbf{S}]\big\} \\ &= L(\mathbf{T}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2).\end{aligned} \tag{2.4}$$

The maximum likelihood solution for $\mathbf{W}$ results in

$$\mathbf{W_{ML}} = \mathbf{U_M}(\mathbf{L}_M - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R},$$

where the columns of $\mathbf{U}_M$ are the eigenvectors for $\mathbf{S}$ and $\mathbf{L}_M$ is a diagonal matrix with the corresponding eigenvalues at the diagonal [20].

The average variance, that is lost per discarded dimension can be calculated with

$$\sigma^2_{ML} = \frac{1}{D-Q}\sum_{i=Q+1}^{D}\lambda_i.$$

If the variance $\sigma^2 \to 0$, this is equal to non-probabilistic PCA.

For the posterior distribution we have

$$p(\mathbf{z}|\mathbf{t}) = \mathcal{N}(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{t} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}).$$

Choosing a suitable value for Q is not always straight forward. While cross-validation can be used to compare and select the reduced dimension, it becomes computationally costly

for large datasets. In the next chapter, we will introduce a method that will automatically infer the projection dimensionality from the dataset.

To project a given point $\mathbf{x}_n$ in the latent space, back to the data space ($\hat{\mathbf{x}}_n$), we need to calculate the posterior mean of $\mathbf{x}$ from

$$\mathbb{E}[\mathbf{x}_n|\mathbf{t}_n] = \mathbf{M}^{-1}\mathbf{W}_{ML}^T(\mathbf{t}_n - \bar{\mathbf{t}}).$$

As an example, we sample datapoints $\mathbf{t}_n$ from a skewed two-dimensional gaussian distribution (samples in blue), and plot the one-dimensional projection values $\mathbf{x}_n$ after projecting them back into the data space.

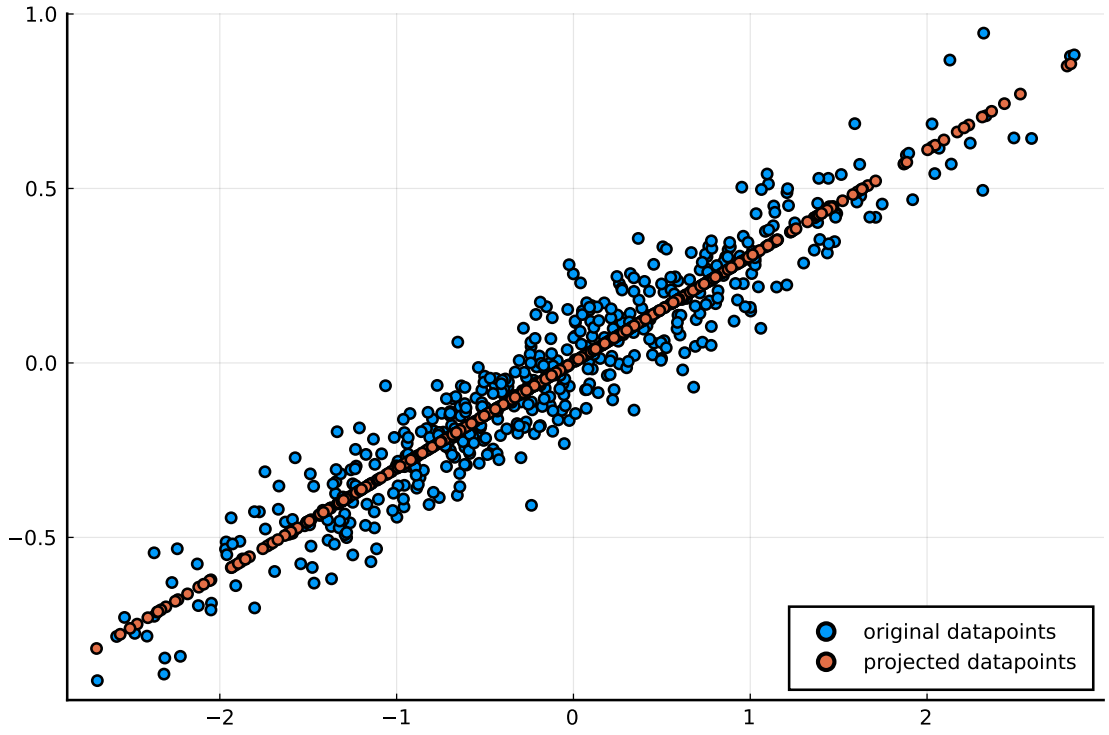

Figure 2.3: PPCA Example. The blue points represent the datapoints $\mathbf{t}_n$, with the orange points represent the projected latent values $\mathbf{x}_n$.

2.3 Bayesian Principal Component Analysis

One drawback with the previous approach of PPCA is, that we need to choose the dimension Q of the principal subspace. While it sometimes is possible to make an educated guess by e.g. plotting the dataset, it is not always easily seen [6].

To derive the dimension of the principal subspace automatically, we make use of the Bayesian framework and its capability to do Automatic Relevance Detection (ARD). Therefore we assign prior distributions to our model parameters and marginalize over them. This method is called BPCA. Because the marginalization is analytically untractable, we will use evidence approximation, which is suitable for a large amount of datapoints and tightly peaked posterior distributions [5].

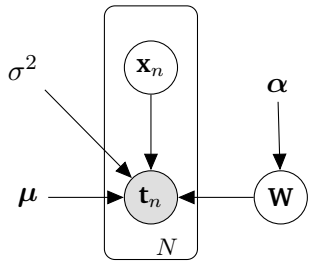

Figure 2.4: Graphical Model of BPCA. Like in PPCA every observation $\mathbf{t}_n$ has a corresponding latent variable $\mathbf{x}_n$. In BPCA, the projection matrix $\mathbf{W}$ is governed by the hyperparameter $\boldsymbol{\alpha}$, to automatically suppress dimensions with low variance.

First we introduce independent gaussian priors over each column of $\mathbf{W}$, where each column represents a basis vector of the principle subspace. We set $\mathbf{W}$ to be an $D \times Q$ matrix, where $Q = D - 1$. While this ensures that the data is simplified by at least one dimension, it does not restrict the solution to use even less dimensions. The resulting dimensionality is defined by the hyperparameters α_i, which control the inverse variance of the corresponding $\mathbf{w}_i$. The probability distribution of $\mathbf{W}$ given $\boldsymbol{\alpha}$ is given by

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{Q} (\frac{\alpha_i}{2\pi})^{D/2} \exp\left\{ -\frac{1}{2}\alpha_i \mathbf{w}_i^T \mathbf{w}_i \right\}. \tag{2.5}$$

By maximizing the marginal likelihood (2.6), some of the α_i in equation (2.5) may be driven to infinity, which in turn will force the corresponding exponent to become increasingly negative and the corresponding $\mathbf{w}_j$ is driven zero, therefore practically eliminating the dimension from the latent space. Using the hyperparameter $\boldsymbol{\alpha}$ to suppress certain $\mathbf{w}_j$ is similar to ARD as introduced by [14]. For the marginal likelihood we define

$$p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2) = \int p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) p(\mathbf{W}|\boldsymbol{\alpha})\, d\mathbf{W}. \tag{2.6}$$

The maximum $\mathbf{W}_{MP}$ can be found by maximizing the log posterior

$$\ln p(\mathbf{W}|\mathcal{D}) = L(\boldsymbol{\mu}, \mathbf{W}, \sigma^2) - \frac{1}{2} \sum_{i=1}^{D-1} \alpha_i \|\mathbf{w}_i\|^2 + const, \tag{2.7}$$

where L is given by (2.4).

When treating $\boldsymbol{\mu}$, σ^2 and $\boldsymbol{\alpha}$ as parameters without prior distributions, we can determine $\boldsymbol{\mu}$ and σ^2 with maximum likelihood, and $\boldsymbol{\alpha}$ using type-II maximum likelihood [5].

As shown by [7], the update for $\boldsymbol{\alpha}$ results in

$$\alpha_i := \frac{\gamma_i}{\|\mathbf{w}_i\|^2},$$

with the effective number of parameters in $\mathbf{w}_i$ given by

$$\gamma_i = d - \alpha_i \operatorname{tr}(\mathbf{H}^{-1}).$$

$\mathbf{H}$ denotes the Hessian matrix, which is the second derivative of $\ln p(\mathbf{W}|\mathcal{D})$ w.r.t $\mathbf{W}$, evaluated at $\mathbf{W}_{MP}$.

To circumvent the calculation of the Hessian matrix, we make the assumption that $\gamma_i = D$, which simplifies the update of $\boldsymbol{\alpha}$ to

$$\alpha_i^{new} = \frac{D}{\|\mathbf{w}_i\|^2}. \tag{2.8}$$

For the derivation of the update for $\mathbf{W}$ we can use the EM algorithm, which is explained in Section 2.4.

In the E-step we evaluate the expected sufficient statistics of the latent-space posterior [5]

$$\mathbb{E}[\mathbf{x}_n] = \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{t}_n - \boldsymbol{\mu}),$$

and

$$\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^T] = \sigma^2\mathbf{M} + \mathbb{E}[\mathbf{x}_n]\,\mathbb{E}[\mathbf{x}_n]^T,$$

with $\mathbf{M}$ given by (2.3).

In the M-step we keep $\mathbf{x}$ constant, while updating $\mathbf{W}$ and σ^2. Getting

$$\mathbf{W}_{new} = \left[\sum_{n=1}^{N}(\mathbf{t}_n - \bar{\mathbf{t}})\,\mathbb{E}[\mathbf{x}_n]^T\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^T] + \sigma^2\mathsf{diag}(\boldsymbol{\alpha})\right]^{-1},$$

and

$$\sigma^2_{new} = \frac{1}{Nd}\sum_{n=1}^{D}\left\{\|\mathbf{t}_n - \boldsymbol{\mu}\|^2 - 2\,\mathbb{E}[\mathbf{x}_n^T]\mathbf{W}_{new}^T(\mathbf{t}_n - \boldsymbol{\mu}) + Tr\big[\,\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^T]\mathbf{W}_{new}^T\mathbf{W}_{new}\big]\right\}.$$

The update of $\mathbf{W}$ and σ^2 is alternated with re-estimation of $\boldsymbol{\alpha}$ using (2.8).

2.3.1 Derivation of Variational Inference Updates

Instead of estimating the parameters $\boldsymbol{\mu}$, σ^2 and $\boldsymbol{\alpha}$ with the maximum likelihood approach, we will now use a fully bayesian treatment and introduce prior distributions for each. The update equations over all parameters are derived by using VI [3, 6, 16].

First we introduce the complete set of model densities

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}_q), \\
p(\mathbf{t}|\mathbf{x}) &= \mathcal{N}(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \tau^{-1}\mathbf{I}_d), \\
p(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu}|\mathbf{0}, \beta^{-1}\mathbf{I}_d), \\
p(\boldsymbol{\alpha}) &= \prod^{q} \Gamma(\alpha_i|a_\alpha, b_\alpha), \\
p(\boldsymbol{\tau}) &= \Gamma(\tau_i|a_\tau, b_\tau), \\
p(\mathbf{W}|\boldsymbol{\alpha}) &= \prod_{i=1}^{q} \left(\frac{\alpha_i}{2\pi}\right)^{\frac{d}{2}} \exp\left\{ -\frac{1}{2}\alpha_i \|\mathbf{w}_i\|^2 \right\}, \\
p(\mathbf{t}) &= \int p(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}),
\end{aligned}
$$

with $\mathbf{C}$ being defined as $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \tau\mathbf{I}$.

To calculate the posterior update equations, we factorize the joint distribution into a posterior distribution q for every parameter like

$$
q(\mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \tau) = q(\mathbf{W}) q(\boldsymbol{\alpha}) q(\boldsymbol{\mu}) q(\tau).
$$

For each function $q(p)$, where p is the corresponding variable, we take the logarithm of the joint density, omitting all terms that do not depend on p, and take the expectations over all variables, except p itself.

2.3.2 Updating $\boldsymbol{\alpha}$

The update equations for the posterior over $\boldsymbol{\alpha}$ can be found by

$$\begin{aligned}
\ln q(\boldsymbol{\alpha}) &= \iiint_{\mathbf{w}\boldsymbol{\mu}\tau} q(\mathbf{w})q(\boldsymbol{\mu})q(\tau)\Big[\ln p(\mathbf{t}|\mathbf{x},\mathbf{W},\boldsymbol{\mu},\tau) + \ln p(\mathbf{W}|\boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha}) + \ln p(\boldsymbol{\mu}) + \ln p(\tau)\Big] \\
&= \int_{\mathbf{w}} q(\mathbf{w})\Big[\ln p(\mathbf{W}|\boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha})\Big] \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \ln \Big[p(\mathbf{W}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})\Big] \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \ln \left[\prod_{i=1}^{q}\left(\frac{\alpha_i}{2\pi}\right)^{\frac{d}{2}} \exp\left\{-\frac{1}{2}\alpha_i\|\mathbf{w}_i\|^2\right\}\prod^{q}\frac{b^a\alpha_i^{a-1}}{\Gamma(a)}e^{-b\alpha_i}\right] \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \ln \left[\prod_{i=1}^{q}\frac{b^a\alpha_i^{a-1}}{\Gamma(a)}\left(\frac{\alpha_i}{2\pi}\right)^{\frac{d}{2}} \exp\left\{-\frac{1}{2}\alpha_i\|\mathbf{w}_i\|^2 - b_\alpha\alpha_i\right\}\right] \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \ln \left[\prod_{i=1}^{q}\frac{b^a}{(2\pi)^{d/2}\Gamma(a)}\alpha_i^{a+\frac{d}{2}-1} \exp\left\{-\left(\frac{1}{2}\|\mathbf{w}_i\|^2 + b\right)\alpha_i\right\}\right] \\
&= \int_{\mathbf{w}} q(\mathbf{w}) \ln \left[\prod_{i=1}^{q}\theta\alpha_i^{\tilde{a}-1}e^{-\tilde{b}(\mathbf{w}_i)\alpha_i}\right].
\end{aligned}$$

We see that the posterior $q(\boldsymbol{\alpha})$ has the form of a Gamma function $\Gamma(\boldsymbol{\alpha}|a_\alpha, b_\alpha)$ with the parameters

$$\begin{aligned}
\tilde{a} &= a + \frac{d}{2}, \\
\tilde{b}(\mathbf{w}_i) &= b + \frac{1}{2}\,\mathbb{E}[\|\mathbf{w}_i\|^2],
\end{aligned}$$

as well as the shape parameter

$$\theta = \frac{b^a}{(2\pi)^{d/2}\Gamma(a)}.$$

2.3.3 Updating $\boldsymbol{\mu}$

For $\boldsymbol{\mu}$ we get

$$\begin{aligned}
\ln q(\boldsymbol{\mu}) &= \mathbb{E}_{W,x,\tau,\alpha}\left[\sum_{n=1}^{N} \ln p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{W}, \boldsymbol{\alpha}, \tau) + \ln p(\boldsymbol{\mu})\right] \\
&= \mathbb{E}_{W,x,\tau,\alpha}\left[\sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \tau^{-1}) + \mathcal{N}(\boldsymbol{\mu}|\mathbf{0}, \beta^{-1})\right] \\
&= \mathbb{E}_{W,x,\tau,\alpha}\left[\sum_{n=1}^{N} \left\{\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\frac{1}{\tau} - \frac{1}{2}\big(\mathbf{t}_n - \mathbf{W}\mathbf{x}_n)^\top \tau \mathbf{I}(\mathbf{t}_n - \mathbf{W}\mathbf{x}_n)\big)\right\}\right. \\
&\qquad \left. + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\frac{1}{\beta} - \frac{1}{2}\boldsymbol{\mu}^\top \beta \mathbf{I}\boldsymbol{\mu}\right] \\
&= \mathbb{E}_{W,x,\tau,\alpha}\left[\sum_{n=1}^{N} \left\{\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\frac{1}{\tau}\right.\right. \\
&\qquad \left.\left. - \frac{1}{2}\big(\mathbf{t}_n^\top \tau \mathbf{t}_n - \mathbf{t}_n^\top \tau \mathbf{W}\mathbf{x}_n - \mathbf{t}_n \tau \boldsymbol{\mu} - \mathbf{x}_n^\top \mathbf{W}^\top \tau \mathbf{t}_n - \mathbf{x}_n^\top \mathbf{W}^\top \tau \mathbf{W}\mathbf{x}_n + \mathbf{x}_n^\top \mathbf{W}^\top \tau \boldsymbol{\mu} + \boldsymbol{\mu}^\top \tau \boldsymbol{\mu}\big)\right\}\right. \\
&\qquad \left. + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\frac{1}{\beta} - \frac{1}{2}\boldsymbol{\mu}^\top \beta \mathbf{I}\boldsymbol{\mu}\right].
\end{aligned}$$

Joining the two quadratic terms $\boldsymbol{\mu}^\top \boldsymbol{\mu}$

$$\left[\sum^{N} \boldsymbol{\mu}^\top \tau \mathbf{I}\boldsymbol{\mu}\right] + \boldsymbol{\mu}^\top \beta \mathbf{I}\boldsymbol{\mu} = \boldsymbol{\mu}^\top (\beta + N\tau)\mathbf{I}\boldsymbol{\mu},$$

we obtain a Normal distribution with variance

$$\boldsymbol{\Sigma}_{\boldsymbol{\mu}} = (\beta + N\overline{\tau})^{-1}\mathbf{I}.$$

Since $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ is a diagonal matrix, it is true that $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^\top = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}$, which leads to the mean

$$\mathbf{m}_{\boldsymbol{\mu}} = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}\overline{\tau}\sum^{N}(\mathbf{t}_n - \overline{\mathbf{W}}\overline{\mathbf{x}}_n).$$

2.3.4 Updating x

The posterior parameters of $\mathbf{x}$ can be found with

$$\begin{aligned}
\ln q(\mathbf{x}_n) &= \mathbb{E}_{\mathbf{W},\boldsymbol{\mu},\tau,\boldsymbol{\alpha}}\left[\ln p(\mathbf{t}_n|\mathbf{W}\mathbf{x}_n+\boldsymbol{\mu},\tau^{-1}+\ln p(\mathbf{W}|\mathbf{0},\boldsymbol{\alpha}^{-1})+\ln p(\tau)+\ln p(\boldsymbol{\alpha})+\ln p(\mathbf{x}_n|\mathbf{0},\mathbf{I})\right] \\
&= \mathbb{E}_{\mathbf{W},\boldsymbol{\mu},\tau}\left[\ln p(\mathbf{t}_n|\mathbf{W}\mathbf{x}_n+\boldsymbol{\mu},\tau^{-1})+\ln p(\mathbf{x}_n|\mathbf{0},\mathbf{I})\right]+C_0 \\
&= \mathbb{E}_{\mathbf{W},\boldsymbol{\mu}}\left[-\frac{1}{2}(\mathbf{t}_n-\mathbf{W}\mathbf{x}_n-\boldsymbol{\mu})^\top\overline{\tau}(\mathbf{t}_n-\mathbf{W}\mathbf{x}_n-\boldsymbol{\mu})+\ln p(\mathbf{x}_0|\mathbf{0},\mathbf{I})+C_1\right] \\
&= \mathbb{E}_{\mathbf{W}}\left[-\frac{1}{2}(\mathbf{t}_n-\mathbf{W}\mathbf{x}_n-\overline{\boldsymbol{\mu}})^\top\overline{\tau}(\mathbf{t}_n-\mathbf{W}\mathbf{x}_n-\overline{\mu})+\ln p(\mathbf{x}_n|\mathbf{0},\mathbf{I})+C_2\right] \\
&= -\frac{1}{2}(\mathbf{t}_n-\overline{\mathbf{W}}\mathbf{x}_n-\overline{\boldsymbol{\mu}})^\top\overline{\tau}(\mathbf{t}_n-\overline{\mathbf{W}}\mathbf{x}_n-\overline{\mu})+\ln p(\mathbf{x}_n|\mathbf{0},\mathbf{I})+C_3 \\
&= -\frac{1}{2}(\mathbf{t}_n-\overline{\mathbf{W}}\mathbf{x}_n-\overline{\boldsymbol{\mu}})^\top\overline{\tau}(\mathbf{t}_n-\overline{\mathbf{W}}\mathbf{x}_n-\overline{\mu})-\frac{1}{2}\mathbf{x}_n^\top\mathbf{x}_n+C_3 \\
&= -\frac{1}{2}\Big[\mathbf{t}_n^\top\overline{\tau}\mathbf{t}_n-\mathbf{t}_n^\top\overline{\tau}\overline{\mathbf{W}}\mathbf{x}_n-\mathbf{t}_n^\top\overline{\tau}\overline{\boldsymbol{\mu}}-\mathbf{x}_n^\top\mathbf{W}^\top\overline{\tau}\mathbf{t}_n+\mathbf{x}_n^\top\overline{\mathbf{W}}^\top\overline{\tau}\overline{\mathbf{W}}\mathbf{x}_n \\
&\quad+\mathbf{x}^\top\overline{\mathbf{W}}^\top\overline{\tau}\overline{\boldsymbol{\mu}}-\overline{\boldsymbol{\mu}}^\top\overline{\tau}\mathbf{t}_n+\overline{\boldsymbol{\mu}}^\top\overline{\tau}\overline{\mathbf{W}}\mathbf{x}_n+\overline{\boldsymbol{\mu}}^\top\overline{\tau}\overline{\boldsymbol{\mu}}+\mathbf{x}_n^\top\mathbf{x}_n\Big]+C_3 \\
&= -\frac{1}{2}\Big[\mathbf{x}_n^\top(\overline{\mathbf{W}}^\top\overline{\tau}\overline{\mathbf{W}}+\mathbf{I})\mathbf{x}_n-(2\mathbf{t}_n^\top\overline{\tau}\overline{\mathbf{W}}-2\overline{\boldsymbol{\mu}}^\top\overline{\tau}\overline{\mathbf{W}})\mathbf{x}_n+...\Big]+C_3 \\
&= -\frac{1}{2}\Big[\mathbf{x}_n^\top\boldsymbol{\Sigma}_\mathbf{x}^{-1}\mathbf{x}_n-2\overline{\tau}(\mathbf{t}_n^\top\overline{\mathbf{W}}-\overline{\boldsymbol{\mu}}^\top\overline{\mathbf{W}})\mathbf{x}_n+...\Big]+C_3.
\end{aligned}$$

The posterior $q(\mathbf{x})$ is given by $q(\mathbf{x})=\mathcal{N}(\mathbf{x}|\mathbf{m}_x,\boldsymbol{\Sigma}_x)$, with

$$\boldsymbol{\Sigma}_\mathbf{x}=(\overline{\mathbf{W}}^\top\overline{\tau}\overline{\mathbf{W}}+\mathbf{I})^{-1},$$

and

$$\mathbf{m}_\mathbf{x}=\overline{\tau}\boldsymbol{\Sigma}_\mathbf{x}\overline{\mathbf{W}}^\top(\mathbf{t}_n-\overline{\boldsymbol{\mu}}).$$

2.3.5 Updating W

For the posterior of $\mathbf{W}$ we can derive

$$
\begin{aligned}
\ln q(\mathbf{w}_j) &= \mathbb{E}_{\boldsymbol{\mu},\tau,\boldsymbol{\alpha},\mathbf{X}}\left[\left[\sum_{n=1}^{N}\ln p(\mathbf{t}|\mathbf{x},\mathbf{W},\boldsymbol{\mu},\tau)\right]+\ln p(\mathbf{x})+\ln p(\mathbf{W}|\boldsymbol{\alpha})+\ln p(\boldsymbol{\alpha})+\ln p(\boldsymbol{\mu})+\ln p(\tau)\right] \\
&= \mathbb{E}_{\boldsymbol{\mu},\tau,\boldsymbol{\alpha},\mathbf{X}}\left[\sum_{n=1}^{N}\left\{\ln p(\mathbf{t}_n|\mathbf{x}_n,\mathbf{w}_j,\boldsymbol{\alpha}_j,\boldsymbol{\mu},\tau)\right\}+\ln p(\mathbf{w}_j|\boldsymbol{\alpha}_j)\right] \\
&= \mathbb{E}_{\boldsymbol{\mu},\tau,\boldsymbol{\alpha},\mathbf{X}}\left[\sum_{n=1}^{N}\left\{\ln p(\mathbf{t}_n|\mathbf{w}_j\mathbf{x}_n+\boldsymbol{\mu},\tau^{-1}\mathbf{I})\right\}+\ln p(\mathbf{w}_j|\boldsymbol{\alpha}_j)\right] \\
&= \mathbb{E}_{\boldsymbol{\mu},\tau,\boldsymbol{\alpha},\mathbf{X}}\left[\sum_{n=1}^{N}\left\{\frac{d}{2}\ln 2\pi+\frac{1}{2}\ln\tau-\frac{1}{2}\big((\mathbf{t}_n-\mathbf{w}_j\mathbf{x}_n-\boldsymbol{\mu})^\top\tau\mathbf{I}(\mathbf{t}_n-\mathbf{w}_j\mathbf{x}_n-\boldsymbol{\mu})\big)\right\}\right. \\
&\qquad \left.+\frac{d}{2}\ln\boldsymbol{\alpha}_j-\frac{d}{2}\ln 2\pi-\frac{1}{2}\boldsymbol{\alpha}_j\|\mathbf{w}_j\|^2\right] \\
&= \mathbb{E}_{\boldsymbol{\mu},\tau,\boldsymbol{\alpha}}\left[\sum_{n=1}^{N}\left\{\frac{d}{2}\ln 2\pi+\frac{1}{2}\ln\tau-\frac{1}{2}\big((\mathbf{t}_n-\mathbf{w}_j\bar{\mathbf{x}}_n-\boldsymbol{\mu})^\top\tau\mathbf{I}(\mathbf{t}_n-\mathbf{w}_j\bar{\mathbf{x}}_n-\boldsymbol{\mu})\big)\right\}\right. \\
&\qquad \left.+\frac{d}{2}\ln\boldsymbol{\alpha}_j-\frac{d}{2}\ln 2\pi-\frac{1}{2}\boldsymbol{\alpha}_j\|\mathbf{w}_j\|^2\right] \\
&= \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\alpha}}\left[\sum_{n=1}^{N}\left\{\frac{d}{2}\ln 2\pi+\frac{1}{2}\ln\overline{\tau}-\frac{1}{2}\big((\mathbf{t}_n-\mathbf{w}_j\bar{\mathbf{x}}_n-\boldsymbol{\mu})^\top\overline{\tau}\mathbf{I}(\mathbf{t}_n-\mathbf{w}_j\bar{\mathbf{x}}_n-\boldsymbol{\mu})\big)\right\}\right. \\
&\qquad \left.+\frac{d}{2}\ln\boldsymbol{\alpha}_j-\frac{d}{2}\ln 2\pi-\frac{1}{2}\boldsymbol{\alpha}_j\|\mathbf{w}_j\|^2\right] \\
&= \mathbb{E}_{\boldsymbol{\alpha}}\left[\sum_{n=1}^{N}\left\{\frac{d}{2}\ln 2\pi+\frac{1}{2}\ln\overline{\tau}-\frac{1}{2}\big((\mathbf{t}_n-\mathbf{w}_j\bar{\mathbf{x}}_n-\overline{\boldsymbol{\mu}})^\top\overline{\tau}\mathbf{I}(\mathbf{t}_n-\mathbf{w}_j\bar{\mathbf{x}}_n-\overline{\boldsymbol{\mu}})\big)\right\}\right. \\
&\qquad \left.+\frac{d}{2}\ln\boldsymbol{\alpha}_j-\frac{d}{2}\ln 2\pi-\frac{1}{2}\boldsymbol{\alpha}_j\|\mathbf{w}_j\|^2\right] \\
&= \sum_{n=1}^{N}\left\{\frac{d}{2}\ln 2\pi+\frac{1}{2}\ln\overline{\tau}-\frac{1}{2}\big((\mathbf{t}_n-\mathbf{w}_j\bar{\mathbf{x}}_n-\overline{\boldsymbol{\mu}})^\top\overline{\tau}\mathbf{I}(\mathbf{t}_n-\mathbf{w}_j\bar{\mathbf{x}}_n-\overline{\boldsymbol{\mu}})\big)\right\} \\
&\quad +\frac{d}{2}\ln\overline{\boldsymbol{\alpha}}_j-\frac{d}{2}\ln 2\pi-\frac{1}{2}\overline{\boldsymbol{\alpha}}_j\|\mathbf{w}_j\|^2
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{n=1}^{N} \Big[\frac{d}{2} \ln 2\pi + \frac{1}{2} \ln \overline{\tau} - \frac{1}{2} \Big(\overline{\mathbf{t}}_n^\top \overline{\tau} \mathbf{t}_n - \mathbf{t}_n^\top \tau \mathbf{w}_j \overline{\mathbf{x}}_n - \mathbf{t}_n^\top \overline{\tau \boldsymbol{\mu}} - \overline{\mathbf{x}}_n^\top \mathbf{w}_j^\top \overline{\tau} \mathbf{t}_n \\
&\qquad + \overline{\mathbf{x}}_n^\top \mathbf{w}_j^\top \overline{\tau} \mathbf{w}_j \overline{\mathbf{x}}_n + \overline{\mathbf{x}}_n^\top \overline{\tau \boldsymbol{\mu}} - \overline{\boldsymbol{\mu}}^\top \overline{\tau} \mathbf{t}_n + \overline{\boldsymbol{\mu}}^\top \overline{\tau} \mathbf{w}_j \overline{\mathbf{x}}_n \Big) \Big] + \frac{d}{2} \ln \overline{\boldsymbol{\alpha}}_j - \frac{d}{2} \ln 2\pi - \frac{1}{2} \overline{\boldsymbol{\alpha}}_j \|\mathbf{w}_j\|^2 \\
&= \frac{Nd}{2} \ln 2\pi + \frac{N}{2} \ln \overline{\tau} - \frac{1}{2} \overline{\tau} \big[\sum_{n=1}^{N} \overline{\mathbf{x}}_n^\top \overline{\mathbf{x}}_n \big] \mathbf{w}_j^\top \mathbf{w}_j - \frac{1}{2} \overline{\boldsymbol{\alpha}}_j \mathbf{w}_j^\top \mathbf{w}_j + \frac{1}{2} \overline{\mathbf{x}}_n^\top \mathbf{w}_j^\top \overline{\tau} (\mathbf{t}_n - \overline{\boldsymbol{\mu}}) \\
&\quad - \frac{1}{2} \overline{\tau} \big[\sum_{n=1}^{N} \mathbf{t}_n^\top \mathbf{t}_n \big] + \frac{1}{2} \tau \sum_{n=1}^{N} \mathbf{t}_n^\top \overline{\boldsymbol{\mu}} + \frac{1}{2} \overline{\boldsymbol{\mu}}^\top \tau \sum_{n=1}^{N} \mathbf{t}_n - \frac{1}{2} \overline{\boldsymbol{\mu}}^\top \overline{\boldsymbol{\mu}} \\
&= \frac{Nd}{2} \ln 2\pi + \frac{N}{2} \ln \overline{\tau} - \frac{1}{2} \big[\overline{\boldsymbol{\alpha}}_j + \overline{\tau} \sum_{n=1}^{N} \overline{\mathbf{x}}_n^\top \overline{\mathbf{x}}_n \big] \mathbf{w}_j^\top \mathbf{w}_j + \frac{1}{2} \overline{\mathbf{x}}_n^\top \mathbf{w}_j^\top \overline{\tau} (\mathbf{t}_n - \overline{\boldsymbol{\mu}}) \\
&\quad - \frac{1}{2} \overline{\tau} \big[\sum_{n=1}^{N} \mathbf{t}_n^\top \mathbf{t}_n \big] + \frac{1}{2} \tau \sum_{n=1}^{N} \mathbf{t}_n^\top \overline{\boldsymbol{\mu}} + \frac{1}{2} \overline{\boldsymbol{\mu}}^\top \tau \sum_{n=1}^{N} \mathbf{t}_n - \frac{1}{2} \overline{\boldsymbol{\mu}}^\top \overline{\boldsymbol{\mu}}.
\end{aligned}
$$

The posterior $q(\mathbf{w}_j)$ is then given by $q(\mathbf{w}_j) = \mathcal{N}(\mathbf{w}_j | \mathbf{m}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$, with

$$
\boldsymbol{\Sigma}_\mathbf{w} = \Big(\overline{\boldsymbol{\alpha}}_j + \overline{\tau} \sum_{n=1}^{N} \overline{\mathbf{x}}_n^\top \overline{\mathbf{x}}_n \Big)^{-1},
$$

and

$$
\mathbf{m}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w} \overline{\tau} \sum_{n=1}^{N} \overline{\mathbf{x}}_n (\mathbf{t}_n - \overline{\boldsymbol{\mu}}_j).
$$

2.3.6 Updating τ

To update the parameters of the posterior over τ we use

$$\begin{aligned}\ln q(\tau) &= \ln \mathbb{E}_{\mathbf{W},\boldsymbol{\alpha},\mathbf{x},\boldsymbol{\mu}} \left[\left[\prod^{N} p(\mathbf{t}_n | \mathbf{x}_n, \mathbf{W}, \boldsymbol{\alpha}, \tau) \right] + \ln p(\tau) \right] \\ &= \frac{N}{2} \ln \tau + (a_0 - 1) \ln \tau - b_0 \tau \\ &\quad - \frac{1}{2} \tau \sum^{N} \left[\mathbf{t}_n^\top \mathbf{t}_n - \mathbf{t}_n^\top \overline{\mathbf{W}} \overline{\mathbf{x}}_n - \mathbf{t}_n^\top \overline{\boldsymbol{\mu}} - \overline{\mathbf{x}}_n^\top \overline{\mathbf{W}}^\top \mathbf{t}_n + \overline{\mathbf{x}}_n^\top \overline{\mathbf{W}}^\top \overline{\mathbf{W}} \overline{\mathbf{x}}_n \right. \\ &\qquad \left. + \overline{\mathbf{x}}_n^\top \overline{\mathbf{W}}^\top \overline{\boldsymbol{\mu}} - \overline{\boldsymbol{\mu}}^\top \mathbf{t}_n + \overline{\boldsymbol{\mu}}^\top \overline{\mathbf{W}} \overline{\mathbf{x}}_n + \overline{\boldsymbol{\mu}}^\top \overline{\boldsymbol{\mu}} \right].\end{aligned}$$

The posterior $q(\tau)$ and the parameter updates result in

$$q(\tau) = \Gamma(\tau | \tilde{a}_\tau, \tilde{b}_\tau),$$

with

$$\begin{aligned}\tilde{a}_\tau &= a_0 + \frac{N}{2}, \text{and} \\ \tilde{b}_\tau &= b_0 + \frac{1}{2} \tau \sum^{N} \left[\mathbf{t}_n^\top \mathbf{t}_n - \mathbf{t}_n^\top \overline{\mathbf{W}} \overline{\mathbf{x}}_n - \mathbf{t}_n^\top \overline{\boldsymbol{\mu}} - \overline{\mathbf{x}}_n^\top \overline{\mathbf{W}}^\top \mathbf{t}_n + \overline{\mathbf{x}}_n^\top \overline{\mathbf{W}}^\top \overline{\mathbf{W}} \overline{\mathbf{x}}_n \right. \\ &\qquad \left. + \overline{\mathbf{x}}_n^\top \overline{\mathbf{W}}^\top \overline{\boldsymbol{\mu}} - \overline{\boldsymbol{\mu}}^\top \mathbf{t}_n + \overline{\boldsymbol{\mu}}^\top \overline{\mathbf{W}} \overline{\mathbf{x}}_n + \overline{\boldsymbol{\mu}}^\top \overline{\boldsymbol{\mu}} \right].\end{aligned}$$

2.3.7 Example Task

Using the stated model and the derived parameter updates, we can now apply BPCA to an example dataset, sampled from a 10-dimensional multivariate normal distribution. We use a Hinton Plot, to visualize the entries of the projection matrix $\mathbf{W}$. Each value is represented by a rectangle, where the size represents the absolute value and the color the sign. The approximation with PCA can be seen in Figure 2.5, and the resulting BPCA approximation of $\mathbf{W}$ in Figure 2.6. We can see that BPCA suppresses the three dimensions with the least variance in the latent space, simplifying the effective dimensionality of the data, by driving the correlating columns $\mathbf{w}_i$ to 0. As an alternative to the hinton plot, we plot the lengths of the column vectors $\mathbf{w}_i$ of the BPCA solution, over the column number i in Figure 2.7.

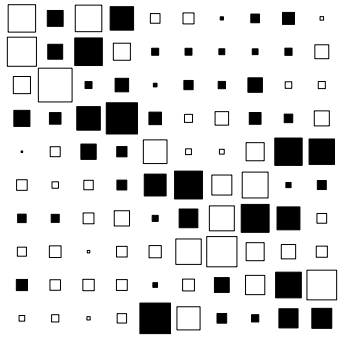

Figure 2.5: PCA Hinton Plot

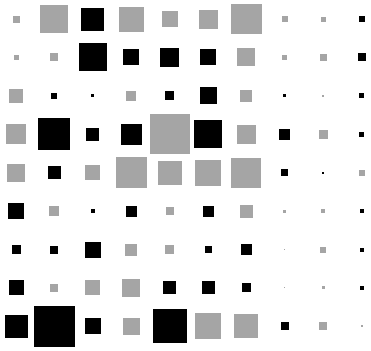

Figure 2.6: BPCA Hinton Plot of matrix $\mathbf{W}$

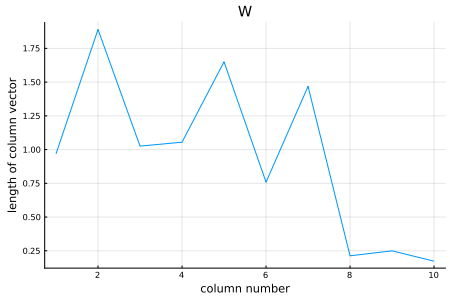

Figure 2.7: Lengths of column vectors in matrix $\mathbf{W}$

2.4 Expectation Maximization

The Expectation Maximization algorithm, is a widely used iterative algorithm, that finds maximum likelihood solutions for probabilistic latent models [6].

Let $\mathbf{X}$ be the observed variables, $\mathbf{Z}$ the latent variables, and $\boldsymbol{\theta}$ be the model parameters.

In general we want to maximize $p(\mathbf{X}|\boldsymbol{\theta})$, but maximizing the complete-data log likelihood $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ is more feasible, so we make use of marginalization. While we use sums here for discrete latent variables, these can be interchanged with integrals when dealing with continuous latent variables. $p(\mathbf{X}|\boldsymbol{\theta})$ is given by

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

A joint probability of $\mathbf{X}$ and $\mathbf{Z}$ can be split into a conditional and a prior part

$$p(\mathbf{X}, \mathbf{Z}) = p(\mathbf{Z}|\mathbf{X})p(\mathbf{X}),$$

which leads to the probability of $\mathbf{X}$ being

$$p(\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}.$$

By taking the logarithm on both sides, and introducing a new function $q(\mathbf{Z})$, we can decompose the right hand side into two parts $\mathcal{L}(q, \boldsymbol{\theta})$ and $\mathrm{KL}(q\|p)$ with

$$\begin{aligned}
\ln p(\mathbf{X}|\boldsymbol{\theta}) &= \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \\
&= \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}}_{\mathcal{L}(q,\boldsymbol{\theta})} \underbrace{- \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}}_{\mathrm{KL}(q\|p)} \\
&= \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p),
\end{aligned}$$

where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}, \tag{2.9}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}. \tag{2.10}$$

Since the Kullback-Leibler divergence (2.9) between q and p, $\mathrm{KL}(q\|p)$ is always greater or equal to 0, we know that $\mathcal{L}(q, \boldsymbol{\theta})$ forms a lower bound on $\ln p(\mathbf{X}|\boldsymbol{\theta})$. This lower bound reaches its maximum when the KL-divergence vanishes, or expressed formally when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.

The idea of the EM algorithm is to maximize this lower bound. The maximization is split into two steps, namely the Expectation Step (E-Step) and the Maximization Step (M-Step). In the E-Step, we keep the parameters $\boldsymbol{\theta}$ fix, while maximizing the lower-bound $\mathcal{L}$ w.r.t $q(\mathbf{Z})$. The obtained new values for $q(\mathbf{Z})$ are now used in the M-Step, where we keep them fixed, and maximize the lower bound $\mathcal{L}$ w.r.t the parameters $\boldsymbol{\theta}$ leading to $\boldsymbol{\theta}^{\text{new}}$. The E- and M-Step are computed iteratively, until the lower bound converges to a local maxima. Instead of maximizing the lower bound in each iteration, it is shown by [6], that it suffices to just increase the lower-bound in each step, leading to the idea of General-EM.

Next the EM algorithm is exemplary applied to a mixture of gaussians.

2.4.1 Mixtures of Gaussians Example

A Gaussian mixture distribution can be written as a linear superposition of Gaussians in the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where π_k are the mixing weights [6].

Because this superposition should result in a probability distribution, the mixture weights π_k have to satisfy $0 \leq \pi_k \leq 1$ and

$$\sum_{k=1}^{K} \pi_k = 1.$$

A K-dimensional binary random variable $\mathbf{z}$ with $z_k \in 0, 1$ and $\sum_k z_k = 1$ is now introduced with $p(z_k = 1) = \pi_k$, and respectively

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

The conditional distribution of the observed variables $\mathbf{x}$ given the latent variable $\mathbf{z}$ can be written as $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and respectively

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

Using the latent variable z_k in the exponent, will ensure that only the belonging distribution is active for each $\mathbf{z}$ and $\mathbf{x}$.

Multiplying the conditional distribution $p(\mathbf{x}|\mathbf{z})$ with $p(\mathbf{z})$ and summing over all $\mathbf{z}$ we obtain the marginal distribution $p(\mathbf{x})$ with

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The posterior probability that a given $\mathbf{x}_n$ belongs to cluster k can be represented by the conditional distribution of $\mathbf{z}$ given $\mathbf{x}$:

$$\begin{aligned} \gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \end{aligned} \tag{2.11}$$

π_k can be seen as the prior probability of $z_k = 1$ and $\gamma(z_k)$ as the posterior probability observing $\mathbf{x}$ or the responsibility that component k takes, for explaining observation $\mathbf{x}$.

Since the posterior probability distributions might overlap, the association of a datapoint with a cluster is given as a probability, making this a soft clustering method.

2.4.2 EM for Gaussian Mixtures Derivation

The joint probability for $\mathbf{x}_n$ and $\mathbf{z}_n$ can be written as

$$\begin{aligned} p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\theta}) &= p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})p(\mathbf{z}_n|\boldsymbol{\theta}) \\ &= \prod_{k=1}^{K} \big[\underbrace{p(\mathbf{x}_n|z_{nk}=1, \boldsymbol{\theta})}_{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \underbrace{p(z_{nk}=1|\boldsymbol{\theta})}_{\pi_k} \big]^{z_{nk}} \\ &= \prod_{k=1}^{K} \big[\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \big]^{z_{nk}}, \end{aligned}$$

and for all datapoints as

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \big[\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \big]^{z_{nk}}.$$

If we now take the logarithm on both sides [6], we obtain

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \big[\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k \boldsymbol{\Sigma}_k) \big]. \tag{2.12}$$

Taking the expectation w.r.t $p(\mathbf{Z}|\mathbf{X}, \theta^i)$ we can calculate the update equations for the E-Step with

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^i)} \big[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \big] = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^i)} \big[z_{nk} \big] \ln \big[\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \big],$$

$$p(z_{nk}=1|\mathbf{x}_n, \boldsymbol{\theta}^i) = \frac{p(\mathbf{x}_n|z_{nk}=1, \boldsymbol{\theta}^i)p(z_{nk}=1|\boldsymbol{\theta}^i)}{\sum_{k^*=1}^{K} p(\mathbf{x}_n|z_{nk^*}=1, \boldsymbol{\theta}^i)p(z_{nk^*}=1|\boldsymbol{\theta}^i)},$$

$$
\begin{aligned}
\mathbb{E}\big[z_{nk}\big] &= 0 \cdot p(z_{nk} = 0|\mathbf{x}_n, \boldsymbol{\theta}^i) \\
&\quad + 1 \cdot p(z_{nk} = 1|\mathbf{x}_n, \boldsymbol{\theta}^i) \\
&= \frac{\pi_k^i \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k^i, \boldsymbol{\Sigma}_k^i)}{\sum_{k^*=1}^{K} \pi_{k^*}^i \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k^*}^i, \boldsymbol{\Sigma}_{k^*}^i)},
\end{aligned} \tag{2.13}
$$

and

$$
\mathbb{E}\big[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\big] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\big[z_{nk}\big]\big[\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k\boldsymbol{\Sigma}_k)\big]. \tag{2.14}
$$

Maximizing (2.13) corresponds to the E-Step, while maximizing (2.14) w.r.t $\boldsymbol{\theta}$ corresponds to the M-Step. In order to maximize (2.14) we need to derive for the several components of $\boldsymbol{\theta}$.

Deriving for $\boldsymbol{\mu}$ only the terms that depend on $\boldsymbol{\mu}$ yields

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_k} \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) =& \frac{\partial}{\partial \boldsymbol{\mu}_k}\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \\
=& \frac{\partial}{\partial \boldsymbol{\mu}_k}\Big(-\frac{1}{2}\big[\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - 2\mathbf{x}_n^\top \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\big]\Big) \\
=& -\frac{1}{2}\big[2\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - 2\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n\big] \\
=& \boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n - \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k.
\end{aligned}
$$

We can now use the equalities

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{x}}\big(\mathbf{x}^\top \mathbf{A}\mathbf{x}\big) &= (\mathbf{A}^\top + \mathbf{A})\mathbf{x}, \\
\frac{\partial}{\partial \mathbf{x}}\big(\mathbf{a}^\top \mathbf{x}\big) &= \mathbf{a},
\end{aligned}
$$

set the derivative of the derivation equal to zero, and solve for the parameter $\boldsymbol{\mu}_k$ with

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathbb{E}\big[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\big] = \sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]\left(\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n - \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \stackrel{!}{=} 0$$
$$\Rightarrow \sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]\mathbf{x}_n = \sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]\boldsymbol{\mu}_k,$$

leading to the update equation for parameter $\boldsymbol{\mu}_k$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]\mathbf{x}_n}{\sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]}.$$

Next we derive the update equation for $\boldsymbol{\Sigma_k}$, and define the precision $\mathbf{A} = \boldsymbol{\Sigma}_k^{-1}$.
Starting with the log normal part of (2.14) and using the equalities

$$\frac{\partial \ln|\mathbf{A}|}{\partial \mathbf{A}} = (\mathbf{A}^{-1})^\top,$$
$$\frac{\partial \operatorname{tr}(\mathbf{AB})}{\partial \mathbf{A}} = \mathbf{B}^\top,$$

we see that

$$\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k\boldsymbol{\Sigma}_k) = \frac{1}{2}\ln|\mathbf{A}_k| - \frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_k)^\top\mathbf{A}_k(\mathbf{x}_n-\boldsymbol{\mu}_k) \tag{2.15}$$
$$= \frac{1}{2}\Big[\ln|\mathbf{A}_k| - \operatorname{tr}\big(\mathbf{A}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^\top\big)\Big]. \tag{2.16}$$

Again we take the derivative of (2.16) w.r.t $\mathbf{A}_k$, resulting in

$$\frac{\partial}{\partial \mathbf{A}_k}\frac{1}{2}\Big[\ln|\mathbf{A}_k| - \operatorname{tr}\big(\mathbf{A}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^\top\big)\Big] \tag{2.17}$$
$$= \frac{1}{2}\Big[\boldsymbol{\Sigma}_k - (\mathbf{x}_n-\boldsymbol{\mu}_k)^\top(\mathbf{x}_n-\boldsymbol{\mu}_k)\Big]. \tag{2.18}$$

Inserting (2.18) into (2.14) and setting the derivative w.r.t $\mathbf{A}$ to zero, we obtain

$$\frac{\partial}{\partial \mathbf{A}_k}\mathbb{E}\big[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\big] \stackrel{!}{=} \mathbf{0}$$

$$\Rightarrow \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]\big[\boldsymbol{\Sigma}_k - (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top(\mathbf{x}_n - \boldsymbol{\mu}_k)\big] \stackrel{!}{=} \mathbf{0}$$

$$\Rightarrow \sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]\boldsymbol{\Sigma}_k = \sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big](\mathbf{x}_n - \boldsymbol{\mu}_k)^\top(\mathbf{x}_n - \boldsymbol{\mu}_k).$$

We can now divide this result by the sum of expectations of z_{nk} and get the result for $\boldsymbol{\Sigma}_k$ with

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big](\mathbf{x}_n - \boldsymbol{\mu}_k)^\top(\mathbf{x}_n - \boldsymbol{\mu}_k)}{\sum_{n^*=1}^{N}\mathbb{E}\big[z_{nk}\big]}.$$

For the derivation of the mixture proportions $\boldsymbol{\pi}$, the constraint $\sum_{k=1}^{K}\pi_k = 1$ needs to be fulfilled. Therefore we use a lagrange multiplier λ in

$$f(x), \text{ subject to } g(x) = 0$$
$$\text{yields } L = f(x) - \lambda g(x),$$

leading to

$$\frac{d}{d\pi_k}\Bigg(\sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]\ln\pi_k - \lambda\big(\sum_{k=1}^{K}\pi_k - 1\big)\Bigg)$$
$$= \frac{\sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big]}{\pi_k} - \lambda \stackrel{!}{=} 0$$
$$=> \sum_{n=1}^{N}\mathbb{E}\big[z_{nk}\big] = \lambda\pi_k$$
$$=> \underbrace{\sum_{n=1}^{N}\underbrace{\sum_{k=1}^{K}\mathbb{E}\big[z_{nk}\big]}_{=1}}_{=N} = \lambda$$
$$=> \lambda = N.$$

The update equation for π_k can be written as

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\big[z_{nk}\big].$$

In Figure 2.8 we generated a dataset from three distinct gaussians, and use EM for clustering.

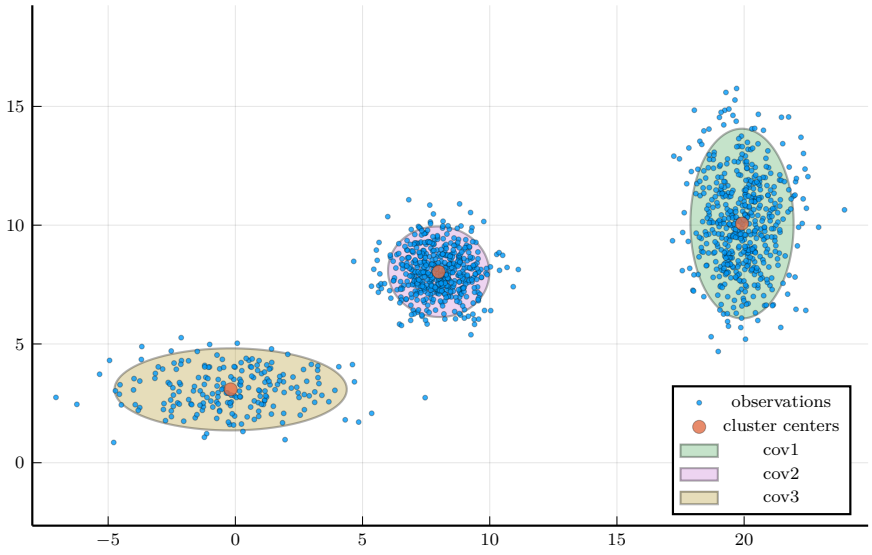

Figure 2.8: Expectation Maximization example – Using a generated dataset, that was sampled from three different multivariate normal distributions, which differ in their mean and covariances. The EM algorithm was able to find and fit three clusters. The cluster centers are depicted in orange, with a different color to represent each fitted covariance.

3 Mixture Models and Probabilistic Principal Component Analysis

We now show that PPCA and BPCA can be easily extended to be used with mixture models. Using a mixture of probabilistic principal component analyzers, we are able to model more complex datasets, while still using simple densities for the mixture components.

3.1 Mixtures of Probabilistic Principal Component Analysers

One big advantage of using a probabilistic formulation for PCA is that we can extent it to a mixture of several analysers. While PCA is limited in its application by its global linearity property, with MPPCA we can combine several local function approximators, enabling better approximations of more complex data structures [19].

In the mixture case, we split the model probability $p(\mathbf{t})$ into several submodels, also called mixture components. Each mixture component $p(\mathbf{t}|k)$ has its own set of parameters $\boldsymbol{\mu}_k$, $\mathbf{W}_k$ and σ_k^2, and is weighted with a mixing proportion π_k, leading to the model probability

$$p(\mathbf{t}) = \sum_{k=1}^{K} \pi_k p(\mathbf{t}|k),$$

where K is the number of components used.

The log likelihood can be formulated accordingly as

$$\begin{aligned} \mathcal{L} &= \sum_{n=1}^{N} \ln \left\{ p(\mathbf{t}_n) \right\} \\ &= \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k p(\mathbf{t}_n|k) \right\}. \end{aligned} \tag{3.1}$$

Using our log likelihood, we can now derive an EM algorithm to get update equations for our parameters. Repeating these updates iteratively will lead to an increasingly precise approximation.

3.1.1 EM for Mixtures of Probabilistic PCA

First we introduce the posterior responsibility R_{nk} of mixture k for generating data point $\mathbf{t}_n$

$$R_{ni} = p(k|\mathbf{t}_n) = \frac{p(\mathbf{t}_n|k)\pi_k}{p(\mathbf{t}_n)}. \tag{3.2}$$

Using the posterior responsibility R_{nk} in the log likelihood equation (3.1), leads to the expected complete-data log likelihood [19]

$$\widehat{\mathcal{L}} = \sum_{n=1}^{N}\sum_{k=1}^{K} R_{nk} \ln\left\{\pi_k p(\mathbf{t}_n|k)\right\}. \tag{3.3}$$

Maximizing (3.3) w.r.t π_k and $\boldsymbol{\mu}_k$ using a Lagrange multiplier leads to the parameter updates

$$\widetilde{\pi}_k = \frac{1}{N}\sum_{n=1}^{N} R_{nk}, \tag{3.4}$$

$$\widetilde{\boldsymbol{\mu}}_k = \frac{\sum_{n=1}^{N} R_{nk}\mathbf{t}_n}{\sum_{n=1}^{N} R_{nk}}. \tag{3.5}$$

To find the update equations for $\mathbf{W}_k$ and σ_k^2, we need to make sure that $\widehat{\mathcal{L}}_C$ is increased with every iteration. While this can be done by maximizing the log likelihood on every step, it is already sufficient to increase it gradually. This is described as Generalized Expectation Maximization (GEM) [6].

Inserting the parameter updates (3.4) and (3.5) into the complete-data log likelihood (3.3) yields

$$\begin{aligned}\left\langle \mathcal{L}_C \right\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K} R_{nk}\Bigg\{ & \ln\widetilde{\pi}_k - \frac{d}{2}\ln\sigma_k^2 - \frac{1}{2}\operatorname{tr}(\left\langle \mathbf{x}_{nk}\mathbf{x}_{nk}^\top \right\rangle) \\ & - \frac{1}{2\sigma_k^2}\|\mathbf{t}_{nk} - \widetilde{\boldsymbol{\mu}}_k\|^2 + \frac{1}{\sigma_k^2}\left\langle \mathbf{x}_{nk}^\top \right\rangle \mathbf{W}_k^\top(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_k) \\ & - \frac{1}{2\sigma_k^2}\operatorname{tr}(\mathbf{W}_k^\top\mathbf{W}_k\left\langle \mathbf{x}_{nk}\mathbf{x}_{nk}^\top \right\rangle)\Bigg\}.\end{aligned} \tag{3.6}$$

Maximizing (3.6) w.r.t to $\mathbf{W}_k$ and σ_k^2, while keeping $\widetilde{\boldsymbol{\mu}}_k$ and $\widetilde{\pi}_k$ fixed [19], we obtain

$$\widetilde{\mathbf{W}}_k = \mathbf{S}_k \mathbf{W}_k (\sigma_k^2 \mathbf{I} + \mathbf{M}_k^{-1} \mathbf{W}_k^\top \mathbf{S}_k \mathbf{W}_k)^{-1}, \tag{3.7}$$

$$\widetilde{\sigma}_k^2 = \frac{1}{d} \operatorname{tr}(\mathbf{S}_k - \mathbf{S}_k \mathbf{W}_k \mathbf{M}_k^{-1} \widetilde{\mathbf{W}}_k^\top), \tag{3.8}$$

where

$$\mathbf{S}_k = \frac{1}{\widetilde{\pi}_k N} \sum_{n=1}^{N} R_{nk} (\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_k)(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_k)^\top.$$

Iterating the parameter updates, alternating between (3.2)-(3.5) and (3.7)-(3.8) guarantees to find a local maximum of the likelihood (3.6).

An alternative way to derive the update equations for $\mathbf{W}_k$ and σ_k^2 is to use eigen-decomposition on $\mathbf{S}_k$ by using

$$\mathbf{W} = \mathbf{U}_q (\boldsymbol{\Lambda}_q - \sigma^2 \mathbf{I})^{1/2} \mathbf{R},$$

$$\sigma_{ML}^2 = \frac{1}{D - Q} \sum_{j=Q+1}^{D} \lambda_j,$$

where the sum from $j = Q + 1$ to D means summing over the smallest $D - Q$ eigenvalues.

However, using the EM approach can have computational advantages [19], since unlike in PPCA, the covariance matrix $\mathbf{S}_k$ for every model has to be re-computed in each iteration.

In Figure 3.1, we use a spiral datasets with added noise and five mixture components to test the algorithm.

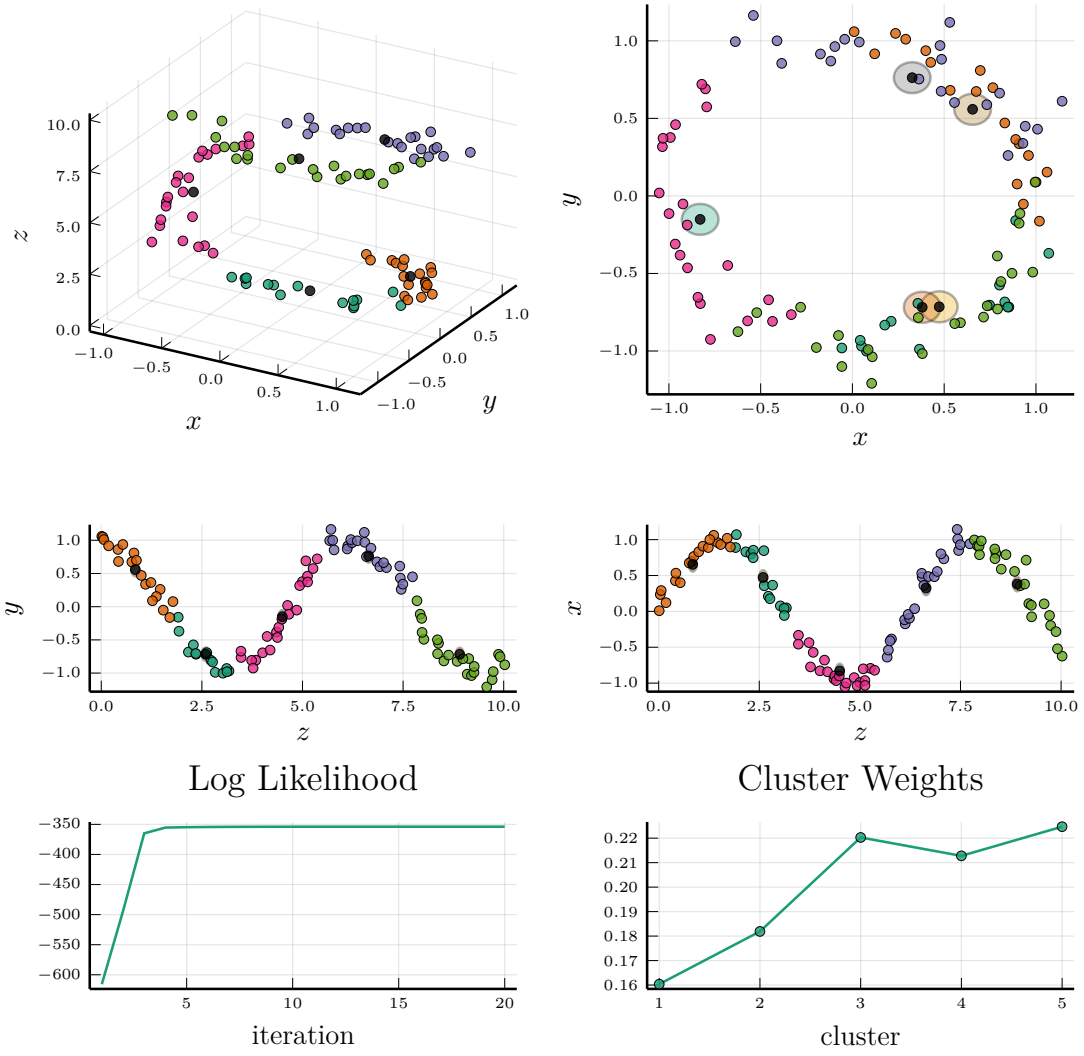

Figure 3.1: MPPCA model with $K = 5$ mixture components, optimized with EM on a spiral dataset with added noise. The black dots represent the means of each component, the datapoints are colored according to their responsible mixture component. At the bottom we see how the EM algorithm monotonically increases the log likelihood function between each iteration and the cluster weights π_k.

3.2 Mixture of Bayesian Principle Component Analysers

After extending PPCA to use mixture models, MPPCA still has the downside of not automatically inferring the dimensionality of the latent space. Furthermore a new hyperparameter, representing the number of models in use, has been introduced, which also needs to be chosen manually. In order to infer these values automatically using Bayesian model comparison [2, 8], we now look at the mixture formulation of BPCA, called BMPCA.

The graphical model from Figure 2.4 of BPCA, is extended as shown in Figure 3.2.

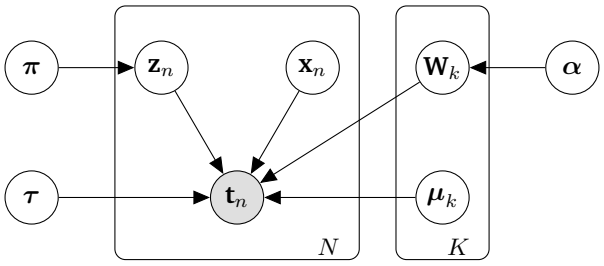

Figure 3.2: Graphical Model of BMPCA – For each of the observations $\mathbf{t}_n$, shaded in gray, there are corresponding latent variables $\mathbf{x}_n$ and $\mathbf{z}_n$. For each model we have a projection matrix $\mathbf{W}_k$, as well as a model mean $\boldsymbol{\mu}_k$. The columns of $\mathbf{W}_k$ depend on the shared hyperparameters $\boldsymbol{\alpha}$.

.

As in MPPCA, we use mixing proportions $\boldsymbol{\pi} = \{\pi_k\}$, and introduce a discrete latent variable $\mathbf{z}_n$ with dimension Q, which is a one-hot vector, for every datapoint $\mathbf{t}_n$, defining which mixture component is responsible for modeling $\mathbf{t}_n$. Instead of having only a single mean $\boldsymbol{\mu}$ and single projection matrix $\mathbf{W}$, we now have one $\boldsymbol{\mu}_k$ and $\mathbf{W}_k$ for each mixture component. We also use the hyperparameter $\boldsymbol{\alpha}$ to automatically prune columns of $\mathbf{W}_k$, which in turn lowers the dimensionality of the latent space. While every sub model could have its own set of $\boldsymbol{\alpha}_k$, we share them between all models, to get a continuous non-linear manifold [8].

The prior distribution for $\mathbf{W}_k$ is given by

$$p(\mathbf{W}_k|\boldsymbol{\alpha}) = \prod_{i=1}^{Q} (\frac{\alpha_i}{2\pi})^{D/2} \exp\left\{ -\frac{1}{2}\alpha_i \mathbf{w}_{ki}^T \mathbf{w}_{ki} \right\},$$

and the latent variables $\mathbf{z}_n$ are modeled by

$$p(\mathbf{z} = \delta_k|\boldsymbol{\pi}) = \pi_k,$$

with δ_k being a Dirac delta, where only the k-th element equals to 1, and all other elements are 0.

For the mixing proportions, we use a Dirichlet prior, paremetrized by $\mathbf{u}$ and normalized by the term $Z(\mathbf{u})$, resulting in

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\mathbf{u}) = \frac{1}{Z(\mathbf{u})} \prod_{k=1}^{K} \pi_k^{u_k - 1} \delta\Big(\sum_{k=1}^{K} \pi_k - 1 \Big).$$

The likelihood of the observation $\mathbf{t}_n$ given the parameters is a Normal distribution with the projected mean $\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k$ and precision τ. The latent variable $\mathbf{x}_n$ and sub-model mean $\boldsymbol{\mu}_k$ are both modeled by Normal distributions with zero mean, and the densities over τ and α_{ki} are Gamma distributions.

The update equations for our parameters, are derived using VI. The likelihood functions and prior distributions for our model are

$$\left[\prod^{N} p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{W}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \tau) \right] p(\mathbf{X})p(\mathbf{Z}\,|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\mathbf{W}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\boldsymbol{\mu})p(\tau). \tag{3.9}$$

We again make the assumption that the $q(...)$-distributions can be factorized. The factorization for this model is given by

$$Q(\mathbf{Z}, \mathbf{X}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \tau) = Q(\mathbf{X}|\,\mathbf{Z})Q(\mathbf{Z})Q(\boldsymbol{\pi})Q(\mathbf{W}|\boldsymbol{\alpha})Q(\boldsymbol{\alpha})Q(\boldsymbol{\mu})Q(\tau). \tag{3.10}$$

3.2.1 Model Definition

The equations describing our model are given by

$$
\begin{aligned}
p(\mathbf{t}_n|\mathbf{W}_k, \boldsymbol{\mu}_k, \tau) &= \mathcal{N}(\mathbf{t}_n|\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k, \tau^{-1}\mathbf{I})^{z_{nk}}, \\
p(\mathbf{W}_k|\boldsymbol{\alpha}) &= \prod_{i=1}^{M}(\frac{\alpha_i}{2\pi})^{D/2}\exp\left\{-\frac{1}{2}\alpha_i\mathbf{w}_{ki}^T\mathbf{w}_{ki}\right\}, \\
p(\alpha_{ki}) &= \Gamma(\alpha_{ki}|a_\alpha, b_\alpha), \\
p(\mathbf{x}_n) &= \mathcal{N}(\mathbf{x}_n|\mathbf{0}, \mathbf{I}), \\
p(\tau) &= \Gamma(\tau|a_\tau, b_\tau), \\
p(\boldsymbol{\mu}_k) &= \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \beta^{-1}\mathbf{I}).
\end{aligned}
$$

3.2.2 Variational Inference Derivation of the Posterior Distribution

We now calculate the log over the posterior distributions of our parameters, starting with α_{ki}.

Updating $\boldsymbol{\alpha}$

To find update equations for the posterior over $\boldsymbol{\alpha}$ we use

$$
\begin{aligned}
\ln q(\alpha_{ki}) &= \mathbb{E}\left[\ln p(\mathbf{w}_{ki}|\alpha_{ki}) + \ln p(\alpha_{ki})\right] \\
&= \mathbb{E}\left[\ln\left((\frac{\alpha_{ki}}{2\pi})^{\frac{d}{2}}\exp\left\{-\frac{1}{2}\alpha_{ki}\|\mathbf{w}_{ki}\|^2\right\}\right) + \ln\Gamma(\alpha_{ki}|a_\alpha, b_\alpha)\right] \\
&= \mathbb{E}\left[\frac{d}{2}\ln\alpha_{ki} - \frac{d}{2}\ln 2\pi - \frac{1}{2}\alpha_{ki}\|\mathbf{w}_{ki}\|^2 - \ln\Gamma(a_\alpha) + a_\alpha\ln b_\alpha + (a_\alpha - 1)\ln\alpha_{ki} - b_\alpha\alpha_{ki}\right] \\
&= \mathbb{E}\left[\ln\alpha_{ki}(\frac{d}{2} + a_\alpha - 1) + \alpha_{ki}(-b_\alpha - \frac{1}{2}\|\mathbf{w}_{ki}\|^2) - \frac{d}{2}\ln 2\pi - \ln\Gamma(a_\alpha) + a_\alpha\ln b_\alpha\right] \\
&= \ln\alpha_{ki}(\frac{d}{2} + a_\alpha - 1) + \alpha_{ki}(-b_\alpha - \frac{1}{2}\,\mathbb{E}_{\mathbf{w}}\left[\|\mathbf{w}_{ki}\|^2\right]) + \text{const.}
\end{aligned}
$$

By taking the expectations over all parameters except α_{ki}, we find a solution, that can be expressed as a Gamma Distribution

$$Q(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \prod_{i=1}^{Q} \Gamma(\alpha_{ki} | \widetilde{a}_\alpha, \widetilde{b}_\alpha^{(ki)}),$$

with

$$\begin{aligned} \widetilde{a}_\alpha &= a_\alpha + \frac{d}{2}, \\ \widetilde{b}_\alpha^{(ki)} &= b_\alpha^{(ki)} + \frac{1}{2} \mathbb{E}_{\mathbf{w}} \left[\|\mathbf{w}_{ki}\|^2 \right]. \end{aligned}$$

Updating x

The parameter updates for $\mathbf{x}$ can be derived with

$$\begin{aligned} \ln q(\mathbf{x}_n | \mathbf{z}_n = \delta_k) &= \mathbb{E} \left[z_{nk} \ln p(\mathbf{t}_n, \mathbf{x}_n, \mathbf{W}_k, \boldsymbol{\mu}_k, \tau) + \ln p(\mathbf{x}_n) \right] \\ &= \mathbb{E} \Bigg[\frac{d}{2} \ln 2\pi + \frac{1}{2} \ln \tau - \frac{1}{2} \tau \big(\mathbf{t}_n^\top \mathbf{t}_n - \mathbf{t}_n^\top \mathbf{W}_k \mathbf{x}_n - \mathbf{t}_n^\top \boldsymbol{\mu}_k - \mathbf{x}_n^\top \mathbf{W}_k^\top \mathbf{t}_n + \mathbf{x}_n^\top \mathbf{W}_k^\top \mathbf{W}_k \mathbf{x}_n \\ &+ \mathbf{x}_n^\top \mathbf{W}_k^\top \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^\top \mathbf{t}_n + \boldsymbol{\mu}_k^\top \mathbf{W}_k \mathbf{x}_n + \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k \big) + \frac{d}{2} \ln 2\pi + \frac{1}{2} \ln |\mathbf{I}| - \frac{1}{2} \mathbf{x}_n^\top \mathbf{x}_n \Bigg] \\ &= \mathbf{x}_n^\top \big(\mathbf{I} + \mathbb{E}_\tau[\tau] \, \mathbb{E}_w[\mathbf{W}_k^\top \mathbf{W}_k] \big) \mathbf{x}_n + \mathbf{x}_n \, \mathbb{E}_\tau[\tau] \Big(\mathbb{E}_w[\mathbf{W}_k^\top] (\mathbf{t}_n - \mathbb{E}_x[\mathbf{x}_n | k]) \Big) + \cdots . \end{aligned}$$

Sorting all terms by their dependency on $\mathbf{x}$, $\mathbf{x}_n^\top \mathbf{x}_n$ and those that are not dependend on $\mathbf{x}$, we can see that the posterior distribution $q(\mathbf{x}_n | \mathbf{z}_n = \delta_k)$ is a Normal distribution, with mean $\mathbf{m}_x$ and covariance $\boldsymbol{\Sigma}_x$. The derivation yields

$$\begin{aligned} Q(\mathbf{X} | \mathbf{Z}) &= \prod^{N} Q(\mathbf{x}_n | \mathbf{z}_n = \delta_k), \\ Q(\mathbf{x}_n | \mathbf{z}_n = \delta_k) &= \mathcal{N}(\mathbf{x}_n | \mathbf{m}_x^{(nk)}, \boldsymbol{\Sigma}_x^{(k)}), \\ \mathbf{m}_x^{(nk)} &= \bar{\tau} \boldsymbol{\Sigma}_x^{(k)} \bar{\mathbf{W}}_k^\top (\mathbf{t}_n - \bar{\boldsymbol{\mu}}_k), \\ \boldsymbol{\Sigma}_x^{(k)} &= \big(\mathbf{I} + \bar{\tau} \, \mathbb{E}_w[\mathbf{W}_k^\top \mathbf{W}_k] \big)^{-1}. \end{aligned}$$

Updating τ

The posterior distribution over τ can be written as

$$\begin{aligned}
\ln q(\tau) &= \mathbb{E}\left[\sum^{N}\sum^{K} \ln p(\mathbf{t}_n|\mathbf{W}_k, \boldsymbol{\mu}_k, \tau) + \ln p(\tau)\right] \\
&= \mathbb{E}\left[\sum^{N}\sum^{K} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k)^{z_{nk}} + \ln p(\tau)\right] \\
&= \mathbb{E}\left[\sum^{N}\sum^{K} z_{nk}\left[\frac{d}{2}\ln 2\pi + \frac{1}{2}\ln\tau - \frac{1}{2}\tau\Big(\mathbf{t}_n^\top\mathbf{t}_n - \mathbf{t}_n^\top\mathbf{W}_k\mathbf{x}_n - \mathbf{t}_n^\top\boldsymbol{\mu}_k - \mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{t}_n + \mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{W}_k\mathbf{x}_n\right. \right.\right. \\
&\quad \left.\left.\left. + \mathbf{x}_n^\top\mathbf{W}_k^\top\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^\top\mathbf{t}_n + \boldsymbol{\mu}_k^\top\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k^\top\boldsymbol{\mu}_k\Big)\right] - \ln\Gamma(a_\tau) + a_\tau\ln b_\tau + (a_\tau - 1)\ln\tau - b_\tau\tau\right] \\
&= \mathbb{E}\left[\ln\tau\big(a_\tau - 1 + \sum^{N}\sum^{K}\frac{1}{2}\big)\right. \\
&\quad + \tau\Bigg(-b_\tau - \frac{1}{2}\sum^{N}\sum^{K} z_{nk}\Big[\mathbf{t}_n^\top\mathbf{t}_n - \mathbf{t}_n^\top\mathbf{W}_k\mathbf{x}_n - \mathbf{t}_n^\top\boldsymbol{\mu}_k - \mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{t}_n + \mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{W}_k\mathbf{x}_n \\
&\quad \left.\left. + \mathbf{x}_n^\top\mathbf{W}_k^\top\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^\top\mathbf{t}_n + \boldsymbol{\mu}_k^\top\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k^\top\boldsymbol{\mu}_k\Big]\right)\right].
\end{aligned}$$

Taking the expectations and sorting the terms leads to

$$\ln\tau\big(\underbrace{a_\tau - 1 + \frac{Nd}{2}}_{\widetilde{a}_\tau - 1}\big),$$

$$\widetilde{a}_\tau = a_\tau - 1 + \frac{Nd}{2},$$

and

$$\begin{aligned}
\widetilde{b}_\tau = b_\tau + \frac{1}{2}\sum^{N}\sum^{K}\mathbb{E}[z_{nk}]\Big[&\mathbf{t}_n^\top\mathbf{t}_n + \boldsymbol{\mu}_k^\top\boldsymbol{\mu}_k + \mathrm{tr}(\mathbb{E}[\mathbf{W}_k^\top\mathbf{W}_k]\,\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^\top|k]) \\
&+ 2\,\mathbb{E}[\boldsymbol{\mu}_k^\top]\,\mathbb{E}[\mathbf{W}_k]\,\mathbb{E}[\mathbf{x}_n|k] - 2\mathbf{t}_n^\top\,\mathbb{E}[\mathbf{W}_k]\,\mathbb{E}[\mathbf{x}_n|k] - 2\mathbf{t}_n^\top\,\mathbb{E}[\boldsymbol{\mu}_k]\Big].
\end{aligned}$$

Updating the projection matrix W

To update the parameters of the posterior over $\mathbf{W}$ we can write

$$\begin{aligned}
\ln q(\mathbf{w}_{ki}) &= \mathbb{E}\left[\sum^{N}\sum^{K} \ln p(\mathbf{t}_n|\mathbf{W}_k, \boldsymbol{\mu}_k, \tau) + \ln p(\mathbf{w}_{ki}|\boldsymbol{\alpha}_{ki})\right] \\
&= \mathbb{E}\left[\sum^{N}\sum^{K} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k)^{z_{nk}} + \ln p(\tau)\right] \\
&= \mathbb{E}\Bigg[\sum^{N}\sum^{K} z_{nk}\Bigg[\frac{d}{2}\ln 2\pi + \frac{1}{2}\ln\tau \\
&\qquad - \frac{1}{2}\tau\Big(\mathbf{t}_n^\top\mathbf{t}_n - \mathbf{t}_n^\top\mathbf{W}_k\mathbf{x}_n - \mathbf{t}_n^\top\boldsymbol{\mu}_k - \mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{t}_n + \mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{W}_k\mathbf{x}_n \\
&\qquad + \mathbf{x}_n^\top\mathbf{W}_k^\top\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^\top\mathbf{t}_n + \boldsymbol{\mu}_k^\top\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k^\top\boldsymbol{\mu}_k\Big)\Bigg] \\
&\quad + \ln\left[\left(\frac{\alpha_k}{2\pi}\right)^{d/2}\exp\left\{-\frac{1}{2}\alpha_k\|\mathbf{w}_{ki}\|^2\right\}\right].
\end{aligned}$$

Sorting the terms and evaluating expectations we get

$$\boldsymbol{\Sigma}_{\mathbf{w}}^{(k)} = \left(\operatorname{diag}\mathbb{E}[\boldsymbol{\alpha}_k] + \mathbb{E}[\tau]\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\,\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^\top|k]\right)^{-1}$$

and

$$\mathbf{m}_{\mathbf{w}}^{(ki)} = \boldsymbol{\Sigma}_{\mathbf{w}}\,\mathbb{E}[\tau]\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\,\mathbb{E}[\mathbf{x}_n|k](t_{ni} - \mathbb{E}[\mu_i]).$$

Updating the means $\boldsymbol{\mu}$

For $\boldsymbol{\mu}$ we can derive the parameter updates with

$$
\begin{aligned}
\ln q(\boldsymbol{\mu}) &= \mathbb{E}\left[\sum^{N}\sum^{K} \ln p(\mathbf{t}_n|\mathbf{W}_k, \boldsymbol{\mu}_k, \tau) + \ln p(\boldsymbol{\mu})\right] \\
&= \mathbb{E}\left[\sum^{N}\sum^{K} z_{nk} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k) + \ln \mathcal{N}(\boldsymbol{\mu}|\mathbf{0}, \beta^{-1}\mathbf{I})\right] \\
&= \mathbb{E}\Bigg[\sum^{N}\sum^{K} z_{nk}\Bigg[\frac{d}{2}\ln 2\pi + \frac{1}{2}\ln\tau \\
&- \frac{1}{2}\tau\Big(\mathbf{t}_n^\top\mathbf{t}_n - \mathbf{t}_n^\top\mathbf{W}_k\mathbf{x}_n - \mathbf{t}_n^\top\boldsymbol{\mu}_k - \mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{t}_n + \mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{W}_k\mathbf{x}_n \\
&+ \mathbf{x}_n^\top\mathbf{W}_k^\top\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^\top\mathbf{t}_n + \boldsymbol{\mu}_k^\top\mathbf{W}_k\mathbf{x}_n + \boldsymbol{\mu}_k^\top\boldsymbol{\mu}_k\Big)\Bigg]\Bigg] \\
&+ \frac{d}{2}\ln 2\pi + \frac{1}{2}\ln|\beta^{-1}\mathbf{I}| - \frac{1}{2}\boldsymbol{\mu}^\top\beta^{-1}\mathbf{I}\boldsymbol{\mu},
\end{aligned}
$$

which can be written as

$$
Q(\boldsymbol{\mu}) = \prod^{K}\mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_{\boldsymbol{\mu}}^{(k)}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(k)}),
$$

where

$$
\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(k)} = \left(\beta + \mathbb{E}[\tau]\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\right)^{-1}\mathbf{I}_d
$$

and

$$
\mathbf{m}_{\boldsymbol{\mu}}^{(k)} = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(k)}\,\mathbb{E}[\tau]\sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{t}_n - \mathbb{E}[\mathbf{W}_k]\,\mathbb{E}[\mathbf{x}_n|k]).
$$

Updating the mixture weights π

The updates for the parameters of the mixture weights $\boldsymbol{\pi}$ can be derived with

$$\begin{aligned}
\ln q(\pi_k) &= \mathbb{E}[\ln \prod^{N} p(z_{nk}|\pi_k)] + \mathbb{E}[\ln p(\pi_k)] \\
&= \mathbb{E}[\sum^{N} z_{nk} \ln \pi_k] + \mathbb{E}[\ln \mathrm{Dir}(\pi_k|\widetilde{u}^{(k)})] \\
&= \mathbb{E}[(\sum^{N} z_{nk}) \ln \pi_k] + \mathbb{E}[(u_k - 1) \ln \pi_k] + \text{const} \\
&= \underbrace{\mathbb{E}[(\sum^{N} z_{nk})] + (u_k - 1)}_{(\widetilde{u}_k - 1)} \ln \overline{\pi}_k + \text{const},
\end{aligned}$$

leading to the update

$$\widetilde{u}_k = u_k + \sum_{n=1}^{N} \mathbb{E}[z_{nk}].$$

To summarize the derivation, the resulting q-distributions conclude to

$$\begin{aligned}
Q(\boldsymbol{\mu}) &= \prod^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_{\boldsymbol{\mu}}^{(k)}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(k)}), \\
Q(\mathbf{W}) &= \prod_{k=1}^{K} \prod_{i=1}^{D} \mathcal{N}(\widetilde{\boldsymbol{w}}_{ki}|\mathbf{m}_{w}^{(ki)}, \boldsymbol{\Sigma}_{w}^{(k)}), \\
Q(\tau) &= \Gamma(\tau|\widetilde{a}_\tau, \widetilde{b}_\tau), \\
Q(\boldsymbol{\Pi}) &= \prod^{K} \mathrm{Dir}(\boldsymbol{\pi}_k|\widetilde{\mathbf{u}}^{(k)}), \\
Q(\mathbf{Z}) &= \prod^{N} Q(\mathbf{z}_n).
\end{aligned}$$

In Figure 3.3, we apply the derived equations on a 3-dimensional spiral dataset and show the distribution of the mixture components along the spiral. Also shown in color, is the assignment of the datapoints to the sub-models, given by z_{nk}. We see that the BMPCA algorithm suppresses some of the mixture components by driving their mean values to 0, and not assigning any datapoints to them.

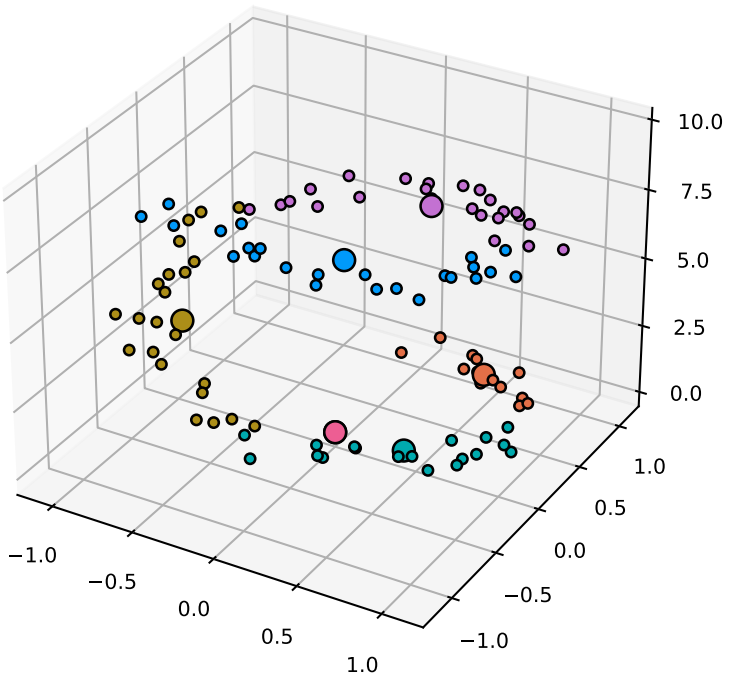

Figure 3.3: BMPCA applied to a three-dimensional spiral dataset, using 7 mixture components. The bigger points describe the mean of each model, while the smaller points represent the given data. Each point is colored according to its assignment to one of the mixture components. In this case, model number 4 and 5 have been automatically suppressed by the BMPCA algorithm, and their means were forced to be 0, moving them to the origin.

4 Variational Locally Projected Regression

After building up the foundations, we now formulate the proposed Variational Locally Projected Regression model.

4.1 Motivation

When doing regression on a dataset, using probabilistic mixture models, often times the data contains several repetitive structures. As a simple example, lets do regression on a sine function, using 6 linear models as the mixture components.

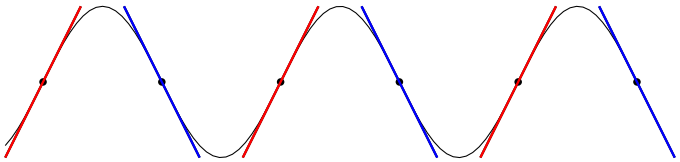

Figure 4.1: Sine Example – Approximation of a sine function, with several linear local models. The red and blue lines illustrate how multiple local models can have similar parameter solutions, when repeating structure is present in a dataset.

The points in Figure 4.1 depict the learned mean of each component, while the lines represent a learned parameter, in this case the slope of the underlying sine function. We can see that all the sub-models shown in red, have the same model parameters, except for a shift in the mean. The same happens with all the sub-models shown in blue.

When doing regression on this data directly like Dirichlet Process Mixtures of Generalized Linear Models (DP-GLM) [17], this repeating structure can result in several mixture models having similar solutions for their parameters. To reduce the parameter redundancy, we project the dataset into a higher-dimensional latent space, such that the repeating structure simplifies under the projection. In the sine-example, the means of each "red" model, and the means of each "blue" model, could be projected onto the same coordinate in the latent space. Doing this will enable us to approximate the dataset with fewer mixture components or fewer parameters.

Let the dataset consist of pairs of datapoints $(\mathbf{x}_n, \mathbf{y}_n)$ which can be seen as the inputs and outputs of the underlying process that we want to approximate. Both $\mathbf{x}_n$ and $\mathbf{y}_n$ are vectors of dimension D. Furthermore let K be the number of mixture components.

We now project each input sample $\mathbf{x}_n$ onto a corresponding latent variable $\mathbf{h}_n$ by using a projection matrix $\mathbf{W}_k$ and a translation vector $\mathbf{b}_k$, such that $\mathbf{h}_n = \mathbf{W}_k\mathbf{x}_n + \mathbf{b}_k$. The latent variable $\mathbf{h}_n$ is of dimension Q, where $Q > D$.

We then transform $\mathbf{h}_n$ into the ouput space with dimension D, projecting it onto the output samples $\mathbf{y}_n$. For this projection we introduce a projection matrix $\mathbf{A}_k$ and a translation vector $\mathbf{c}_k$, such that $\mathbf{y}_n = \mathbf{A}_k\mathbf{h}_n + \mathbf{c}_k$. Which of the components relates to a given datapoint is represented in a latent variable z_{nk}, which is equal to 1 for only one component k per datapoint n, and 0 for all other components.

In order to find a set of parameters, that minimize the approximation error, we use VI to find iterative update equations.

The graphical model of this approach is shown in Figure 4.2.

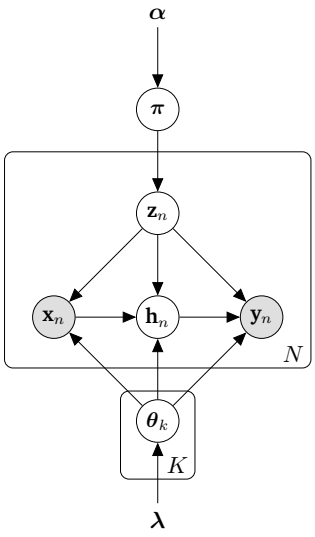

Figure 4.2: The input space $\mathbf{x}_n$ is projected onto the latent variable $\mathbf{h}_n$ and then project onto the output space $\mathbf{y}_n$. The K mixture components are characterized by a set of parameters $\boldsymbol{\theta}_k$ and assigned to each datapoint by $\mathbf{z}_n$.

4.2 Model Definition

We now construct the model equations for Variational Locally Projected Regression. To formulate this within a probabilistic framework, we introduce prior distributions over the model parameters, as well as the Likelihood functions for the variables.

The categorical mixture weights $\boldsymbol{\pi}$ are sampled from a Dirichlet distribution, which is parametrized by a vector $\mathbf{u}$, as

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\pi}|\mathbf{u}).$$

The projection matrices $\mathbf{W}$ and $\mathbf{A}$ are sampled from Matrix-Normal distributions, with their variance matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Delta}$ come from a Wishart distribution, and the translation vectors $\mathbf{b}$ and $\mathbf{c}$ from Normal distributions. The sampling process is described by

$$\begin{aligned}
\boldsymbol{\Lambda} &\sim \mathbf{W}(\boldsymbol{\Psi}, \nu), \\
\boldsymbol{\Delta} &\sim \mathbf{W}(\boldsymbol{\Theta}, \phi), \\
\mathbf{W} &\sim \mathcal{MN}(\mathbf{V}, \boldsymbol{\Lambda}, \mathbf{P}), \\
\mathbf{A} &\sim \mathcal{MN}(\mathbf{M}, \boldsymbol{\Delta}, \mathbf{K}), \\
\mathbf{b} &\sim \mathcal{N}(\mathbf{m}_b, \kappa\boldsymbol{\Lambda}), \\
\mathbf{c} &\sim \mathcal{N}(\mathbf{m}_c, g\boldsymbol{\Delta}).
\end{aligned}$$

The variance matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Delta}$ are shared between the Matrix-Normal and Normal distributions.

We sample the mean and variance for the density over $\mathbf{x}$ from a Normal-Wishart distribution

$$\boldsymbol{\Gamma} \sim \mathcal{W}(\boldsymbol{\Xi}, \boldsymbol{\xi}), \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\gamma}, \beta\boldsymbol{\Gamma}),$$

and to generate $\mathbf{x}$ we use a Normal distribution

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Gamma}).$$

The one-hot labels $\mathbf{z}$ are sampled from a Categorical distribution parametrized by $\boldsymbol{\pi}$ with

$$\mathbf{z} \sim \mathrm{Cat}(\boldsymbol{\pi}).$$

The density of $\mathbf{h}$ given the selection variable $\mathbf{z}$, the input datapoints $\mathbf{x}$ is modeled by $\mathbf{h} \sim \mathcal{N}(\mathbf{W}\mathbf{x} + \mathbf{b}, \mathbf{\Lambda})$, and $\mathbf{y}$ is generated by $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\mathbf{h} + \mathbf{c}, \mathbf{\Lambda})$.

The complete set of model equations with the complete data likelihood is summed up below.

Complete data likelihood

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{h}, \mathbf{z})p(\mathbf{h}|\mathbf{x}, \mathbf{z})p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$$

Likelihood distributions

$$\begin{aligned} p(z = \delta_m|\boldsymbol{\pi}) &= \pi_m \\ p(\mathbf{h}|\mathbf{x}, \mathbf{z}) &= \mathcal{N}(\mathbf{h}|\mathbf{z}, \mathbf{W}\mathbf{x} + \mathbf{b}, \mathbf{\Lambda}) \\ p(\mathbf{y}|\mathbf{h}, \mathbf{z}) &= \mathcal{N}(\mathbf{y}|\mathbf{z}, \mathbf{A}\mathbf{h} + \mathbf{c}, \mathbf{\Delta}) \\ p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x}|\mathbf{z}, \boldsymbol{\eta}, \mathbf{\Gamma}) \end{aligned}$$

Prior distributions

$$\begin{aligned} p(\mathbf{W}, \mathbf{\Lambda}, \mathbf{b}) &= \mathcal{N}(\mathbf{W}|\mathbf{V}, \mathbf{\Lambda}, \mathbf{P})\mathcal{W}(\mathbf{\Lambda}|\mathbf{\Psi}, \nu)\mathcal{N}(\mathbf{b}|\mathbf{m}_b, \kappa\mathbf{\Lambda}) \\ p(\mathbf{A}, \mathbf{\Delta}, \mathbf{c}) &= \mathcal{N}(\mathbf{A}|\mathbf{M}, \mathbf{\Delta}, \mathbf{K})\mathcal{W}(\mathbf{\Delta}|\mathbf{\Theta}, \phi)\mathcal{N}(\mathbf{c}|\mathbf{m}_c, g\mathbf{\Delta}) \\ p(\boldsymbol{\eta}|\mathbf{\Gamma})p(\mathbf{\Gamma}) &= \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{\gamma}_0, \beta\mathbf{\Gamma})\mathcal{W}(\mathbf{\Gamma}|\mathbf{\Xi}_0, \xi_0) \\ p(\boldsymbol{\pi}) &= \mathrm{Dir}(\boldsymbol{\pi}|\mathbf{u}) \end{aligned}$$

4.3 Variational Inference Derivation

We now derive update equations with variational inference for all variables $\mathbf{\Theta}$

$$\mathbf{\Theta} = [\mathbf{z}, \boldsymbol{\pi}, \mathbf{W}, \mathbf{\Lambda}, \mathbf{A}, \mathbf{\Delta}, \boldsymbol{\eta}, \mathbf{\Gamma}, \boldsymbol{h}].$$

Here we use the mean-field assumption, which states that the posterior densities can be factorized with

$$Q(\mathbf{\Theta}) = Q(\mathbf{z})Q(\boldsymbol{\pi})Q(\mathbf{W}, \mathbf{\Lambda})Q(\mathbf{A}, \mathbf{\Delta})Q(\boldsymbol{\eta}, \mathbf{\Gamma})Q(\mathbf{h}).$$

4.3.1 Updating the mixture weights π

To derive the posterior over $\boldsymbol{\pi}$, we take the expectation over all variables except $\boldsymbol{\pi}$, for the log of all terms that depend on $\boldsymbol{\pi}$, using

$$\begin{aligned}
\ln q(\pi_k) &= \mathbb{E}_{q(z)}[\ln \prod^{N} p(z_{nk}|\pi_k)] + \mathbb{E}_{q(z)}[\ln p(\pi_k)] \\
&= \mathbb{E}_{q(z)}[\sum^{N} z_{nk} \ln \pi_k] + \mathbb{E}_{q(z)}[\ln \mathrm{Dir}(\pi_k|\widetilde{u}^{(k)})] \\
&= \mathbb{E}_{q(z)}[(\sum^{N} z_{nk}) \ln \pi_k] + \mathbb{E}_{q(z)}[(u_k - 1) \ln \pi_k] + \text{const} \\
&= \underbrace{\mathbb{E}_q(z)[(\sum^{N} z_{nk})] + (u_k - 1)}_{(\widetilde{u}_k - 1)} \ln \pi_k + \text{const}.
\end{aligned}$$

The update for the parameters $\mathbf{u}$ result in

$$\widetilde{u}_k = u_k + \sum^{N} \mathbb{E}[z_{nk}].$$

4.3.2 Updating the projection matrix W

Here we update the parameters of the density over the projection matrix $\mathbf{W}$, to obtain new values for $\mathbf{P}$ and $\mathbf{V}$. We again take the expectations over all variables except of $\mathbf{W}$ itself. For the updates, we get

$$
\begin{aligned}
\ln q(\mathbf{W}|\mathbf{\Lambda}) &= \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{b},\boldsymbol{h})}\left[\ln p(\mathbf{H}|\mathbf{Z},\mathbf{X},\mathbf{W},\boldsymbol{b},\mathbf{\Lambda}) + \ln p(\mathbf{W}|\mathbf{\Lambda})\right] \\
&= \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{b},\boldsymbol{h})}\left[\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk} \ln \mathcal{N}(\boldsymbol{h}_n|\mathbf{W}_k\boldsymbol{x}_n + \boldsymbol{b}_k, \mathbf{\Lambda})\right] + \sum_{k=1}^{K} \ln \mathcal{MN}(\mathbf{W}_k|\mathbf{V}_0, \mathbf{\Lambda}, \mathbf{P}_0,) \\
&= -\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\left[(\bar{\boldsymbol{h}}_n - \mathbf{W}_k\boldsymbol{x}_n - \bar{\boldsymbol{b}}_k)^\top \mathbf{\Lambda}(\bar{\boldsymbol{h}}_n - \mathbf{W}_k\boldsymbol{x}_n - \bar{\boldsymbol{b}}_k) + \mathrm{tr}(\mathbf{\Lambda}\mathbf{\Sigma}_h^{-1})\right] \\
&\quad -\frac{1}{2}\sum_{k=1}^{K} \mathrm{tr}(\mathbf{P}_0(\mathbf{W}_k - \mathbf{V}_0)^\top \mathbf{\Lambda}(\mathbf{W}_k - \mathbf{V}_0)) + \mathrm{const} \\
&= -\frac{1}{2}\sum_{k=1}^{K} \mathrm{tr}(\mathbf{R}\bar{\mathbf{H}}^\top \mathbf{\Lambda}\bar{\mathbf{H}} - 2\mathbf{R}\bar{\mathbf{H}}^\top \mathbf{\Lambda}\mathbf{W}_k\mathbf{X} - 2\mathbf{R}\bar{\mathbf{H}}^\top \mathbf{\Lambda}\bar{\mathbf{B}}_k \\
&\qquad\qquad + \mathbf{R}\mathbf{X}^\top \mathbf{W}_k^\top \mathbf{\Lambda}\mathbf{W}_k\mathbf{X} + 2\mathbf{R}\mathbf{X}^\top \mathbf{W}_k^\top \mathbf{\Lambda}\bar{\mathbf{B}}_k + \mathbf{R}\bar{\mathbf{B}}_k^\top \mathbf{\Lambda}\bar{\mathbf{B}}_k + \mathbf{R}\mathbf{\Lambda}\sum_{n=1}^{N}\mathbf{\Sigma}_h^{-1}) \\
&\quad -\frac{1}{2}\sum_{k=1}^{K} \mathrm{tr}(\mathbf{P}_0\mathbf{W}_k^\top \mathbf{\Lambda}\mathbf{W}_k - 2\mathbf{P}_0\mathbf{W}^\top \mathbf{\Lambda}\mathbf{V}_0 + \mathbf{P}_0\mathbf{V}_0^\top \mathbf{\Lambda}\mathbf{V}_0) + \mathrm{const} \\
&= -\frac{1}{2}\sum_{k=1}^{K} \mathrm{tr}(-2\mathbf{W}_k^\top \mathbf{\Lambda}(\mathbf{V}_0\mathbf{P}_0 + \bar{\mathbf{H}}\mathbf{R}\mathbf{X}^\top - \bar{\mathbf{B}}_k\mathbf{R}\mathbf{X}^\top) + \mathbf{W}_k^\top \mathbf{\Lambda}\mathbf{W}_k(\mathbf{P}_0 + \mathbf{X}\mathbf{R}\mathbf{X}^\top)) + \mathrm{const},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{P} &= \mathbf{P}_0 + \mathbf{X}\mathbf{R}\mathbf{X}^\top \\
\mathbf{V} &= (\mathbf{V}_0\mathbf{P}_0 + \bar{\mathbf{H}}\mathbf{R}\mathbf{X}^\top - \bar{\mathbf{B}}_k\mathbf{R}\mathbf{X}^\top)(\mathbf{P}_0 + \mathbf{X}\mathbf{R}\mathbf{X}^\top)^{-1}.
\end{aligned}
$$

4.3.3 Updating the translation vector b

The parameter updates for $\mathbf{b}$ are given by

$$\begin{aligned}
\ln q(\boldsymbol{b}|\boldsymbol{\Lambda}) &= \mathbb{E}_{q(\boldsymbol{z},\mathbf{W},\boldsymbol{h})}\left[\ln p(\mathbf{Y}|\mathbf{Z},\mathbf{X},\mathbf{W},\boldsymbol{b},\boldsymbol{\Lambda})\right] + \ln p(\boldsymbol{b}|\boldsymbol{\Lambda}) \\
&= -\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\left[(\bar{\boldsymbol{h}}_n - \bar{\mathbf{W}}_k\boldsymbol{x}_n - \boldsymbol{b}_k)^\top\boldsymbol{\Lambda}(\bar{\boldsymbol{h}}_n - \bar{\mathbf{W}}_k\boldsymbol{x}_n - \boldsymbol{b}_k) + \mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_h^{-1})\right] \\
&\quad -\frac{1}{2}\sum_{k=1}^{K}(\boldsymbol{b}_k - \boldsymbol{m}_{b_0})^\top\kappa_0\boldsymbol{\Lambda}(\boldsymbol{b}_k - \boldsymbol{m}_{b_0}) + \text{const} \\
&= -\frac{1}{2}\left[\sum_{k=1}^{K}(\kappa_0 + \sum_{n=1}^{N} r_{nk})\boldsymbol{b}_k^\top\boldsymbol{\Lambda}\boldsymbol{b}_k - 2\sum_{k=1}^{K}\boldsymbol{b}_k^\top\boldsymbol{\Lambda}(\kappa_0\boldsymbol{m}_{b_0} + \sum_{n=1}^{N} r_{nk}(\bar{\boldsymbol{h}}_n - \bar{\mathbf{W}}_k\boldsymbol{x}_n))\right] + \text{const},
\end{aligned}$$

and

$$\begin{aligned}
\kappa &= \kappa_0 + \sum_{n=1}^{N} r_{nk} \\
\boldsymbol{m}_b &= \frac{\kappa_0\boldsymbol{m}_{b_0} + \sum_{n=1}^{N} r_{nk}(\bar{\boldsymbol{h}}_n - \bar{\mathbf{W}}_k\boldsymbol{x}_n)}{\kappa_0 + \sum_{n=1}^{N} r_{nk}}.
\end{aligned}$$

Where we have used the following identities for the quadratic expectation terms,

$$\begin{aligned}
\mathbb{E}_h(\boldsymbol{h}^\top\boldsymbol{\Lambda}\boldsymbol{h}) &= \bar{\boldsymbol{h}}^\top\boldsymbol{\Lambda}\bar{\boldsymbol{h}} + \mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_h^{-1}), \\
\mathbb{E}_b(\boldsymbol{b}^\top\boldsymbol{\Lambda}\boldsymbol{b}) &= \bar{\boldsymbol{b}}^\top\boldsymbol{\Lambda}\bar{\boldsymbol{b}} + \mathrm{tr}(\boldsymbol{\Lambda}\kappa^{-1}\boldsymbol{\Lambda}^{-1}) \\
&= \bar{\boldsymbol{b}}^\top\boldsymbol{\Lambda}\bar{\boldsymbol{b}} + \kappa^{-1}d, \\
\mathbb{E}_w(\mathbf{W}^\top\boldsymbol{\Lambda}\mathbf{W}) &= \mathbf{P}\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^\top) + \mathbf{V}^\top\boldsymbol{\Lambda}\mathbf{V} \\
&= \mathbf{P}d + \mathbf{V}^\top\boldsymbol{\Lambda}\mathbf{V},
\end{aligned}$$

and the definition of the responsibility variable r_{nk}

$$\mathbb{E}[z_{nk}] = r_{nk}.$$

4.3.4 Updating the precision matrix $\mathbf{\Lambda}$

For $\mathbf{\Lambda}$ we can derive

$$
\begin{aligned}
\ln q(\mathbf{\Lambda}) &= \mathbb{E}_{q(\boldsymbol{z},\mathbf{W},\boldsymbol{b},\boldsymbol{h})}\left[\ln p(\mathbf{H}|\mathbf{Z},\mathbf{X},\mathbf{W},\boldsymbol{b},\mathbf{\Lambda})\right] + \ln p(\mathbf{\Lambda}) + \mathbb{E}_{q(\mathbf{W},\mathbf{c})}\left[\ln p(\mathbf{W}|\mathbf{\Lambda}) + \ln p(\boldsymbol{b}|\mathbf{\Lambda})\right] \\
&= \mathbb{E}_{q(\boldsymbol{z},\mathbf{W},\boldsymbol{b},\boldsymbol{h})}\left[\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk} \ln \mathcal{N}(\boldsymbol{h}_n|\mathbf{W}_k\boldsymbol{x}_n + \boldsymbol{b}_k, \mathbf{\Lambda})\right] + \sum_{k=1}^{K} \ln \mathcal{W}(\mathbf{\Lambda}|\mathbf{\Psi}_0, \nu_0) \\
&\quad + \mathbb{E}_{q(\mathbf{W})}\left[\sum_{k=1}^{K} \ln \mathcal{MN}(\mathbf{W}_k|\mathbf{V}_0, \mathbf{P}_0, \mathbf{\Lambda})\right] + \mathbb{E}_{q(\mathbf{c})}\left[\sum_{k=1}^{K} \ln \mathcal{N}(\boldsymbol{b}|\boldsymbol{m}_{b_0}, \kappa_0\mathbf{\Lambda})\right] \\
&= + \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk} \ln|\mathbf{\Lambda}| - \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\mathbb{E}\left[(\boldsymbol{h}_n - \mathbf{W}_k\boldsymbol{x}_n - \boldsymbol{b}_k)^\top \mathbf{\Lambda}(\boldsymbol{h}_n - \mathbf{W}_k\boldsymbol{x}_n - \boldsymbol{b}_k)\right] \\
&\quad + \frac{m}{2}\sum_{k=1}^{K} \ln|\mathbf{\Lambda}| - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\mathrm{tr}((\mathbf{P}_0(\mathbf{W}_k - \mathbf{V}_0)^\top \mathbf{\Lambda}(\mathbf{W}_k - \mathbf{V}_0))\right] \\
&\quad + \frac{1}{2}\sum_{k=1}^{K} \ln|\mathbf{\Lambda}| - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{b}_k - \boldsymbol{m}_{b_0})^\top(\kappa_0\mathbf{\Lambda})(\boldsymbol{b}_k - \boldsymbol{m}_{b_0})\right] \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}(\nu_0 - d - 1)\ln|\mathbf{\Lambda}| - \frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}(\mathbf{\Psi}_0^{-1}\mathbf{\Lambda}) + \text{const} \\
&= - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\Big[\mathrm{tr}(\mathbf{R}\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H} - 2\mathbf{R}\mathbf{H}^\top\mathbf{\Lambda}\mathbf{W}_k\mathbf{X} - 2\mathbf{R}\mathbf{H}^\top\mathbf{\Lambda}\mathbf{B}_k \\
&\qquad\qquad + \mathbf{R}\mathbf{X}^\top\mathbf{W}_k^\top\mathbf{\Lambda}\mathbf{W}_k\mathbf{X} + 2\mathbf{R}\mathbf{X}^\top\mathbf{W}_k^\top\mathbf{\Lambda}\mathbf{B}_k + \mathbf{R}\mathbf{B}_k^\top\mathbf{\Lambda}\mathbf{B}_k)\Big] \\
&\quad - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\Big[\mathrm{tr}(\mathbf{P}_0\mathbf{W}_k^\top\mathbf{\Lambda}\mathbf{W}_k - 2\mathbf{P}_0\mathbf{W}^\top\mathbf{\Lambda}\mathbf{V}_0 + \mathbf{P}_0\mathbf{V}_0^\top\mathbf{\Lambda}\mathbf{V}_0)\Big] \\
&\quad - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\Big[(\boldsymbol{b} - \boldsymbol{m}_{b_0})^\top(\kappa_0\mathbf{\Lambda})(\boldsymbol{b} - \boldsymbol{m}_{b_0})\Big] \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk} \ln|\mathbf{\Lambda}| + \frac{m}{2}\sum_{k=1}^{K} \ln|\mathbf{\Lambda}| + \frac{1}{2}\sum_{k=1}^{K} \ln|\mathbf{\Lambda}| \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}(\nu_0 - d - 1)\ln|\mathbf{\Lambda}| - \frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}(\mathbf{\Psi}_0^{-1}\mathbf{\Lambda}) + \text{const}
\end{aligned}
$$

$$
\begin{aligned}
= & -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\mathbf{R}\bar{\mathbf{H}}^{\top}\boldsymbol{\Lambda}\bar{\mathbf{H}} - 2\mathbf{R}\bar{\mathbf{H}}^{\top}\boldsymbol{\Lambda}\bar{\mathbf{W}}_k\mathbf{X} - 2\mathbf{R}\bar{\mathbf{H}}^{\top}\boldsymbol{\Lambda}\bar{\mathbf{B}}_k \\
& \qquad + \mathbf{R}\mathbf{X}^{\top}\bar{\mathbf{W}}_k^{\top}\boldsymbol{\Lambda}\bar{\mathbf{W}}_k\mathbf{X} + 2\mathbf{R}\mathbf{X}^{\top}\bar{\mathbf{W}}_k^{\top}\boldsymbol{\Lambda}\bar{\mathbf{B}}_k + \mathbf{R}\bar{\mathbf{B}}_k^{\top}\boldsymbol{\Lambda}\bar{\mathbf{B}}_k + \mathbf{R}\boldsymbol{\Lambda}\sum_{n=1}^{N}\boldsymbol{\Sigma}_h^{-1}) \\
& -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\mathbf{P}_0\bar{\mathbf{W}}_k^{\top}\boldsymbol{\Lambda}\bar{\mathbf{W}}_k - 2\mathbf{P}_0\bar{\mathbf{W}}^{\top}\boldsymbol{\Lambda}\mathbf{V}_0 + \mathbf{P}_0\mathbf{V}_0^{\top}\boldsymbol{\Lambda}\mathbf{V}_0) \\
& -\frac{1}{2}\sum_{k=1}^{K}(\bar{\boldsymbol{b}} - \boldsymbol{m}_{b_0})^{\top}(\kappa_0\boldsymbol{\Lambda})(\bar{\boldsymbol{b}} - \boldsymbol{m}_{b_0}) \\
& +\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln|\boldsymbol{\Lambda}| + \frac{m}{2}\sum_{k=1}^{K}\ln|\boldsymbol{\Lambda}| + \frac{1}{2}\sum_{k=1}^{K}\ln|\boldsymbol{\Lambda}| \\
& +\frac{1}{2}\sum_{k=1}^{K}(\nu_0 - d - 1)\ln|\boldsymbol{\Lambda}| - \frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\boldsymbol{\Psi}_0^{-1}\boldsymbol{\Lambda}) + \text{const} \\
= & -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\mathbf{R}(\bar{\mathbf{H}} - \bar{\mathbf{W}}_k\mathbf{X} - \bar{\mathbf{B}}_k)(\bar{\mathbf{H}} - \bar{\mathbf{W}}_k\mathbf{X} - \bar{\mathbf{B}}_k)^{\top}\boldsymbol{\Lambda} + \mathbf{R}\big(\sum_{n=1}^{N}\boldsymbol{\Sigma}_h^{-1}\big)\boldsymbol{\Lambda} \\
& -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}((\bar{\mathbf{W}}_k - \mathbf{V}_0)\mathbf{P}_0(\bar{\mathbf{W}}_k - \mathbf{V}_0)^{\top}\boldsymbol{\Lambda}) \\
& -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\kappa_0(\bar{\boldsymbol{b}} - \boldsymbol{m}_{b_0})(\bar{\boldsymbol{b}} - \boldsymbol{m}_{b_0})^{\top}\boldsymbol{\Lambda}) \\
& +\frac{1}{2}\sum_{k=1}^{K}(\nu_0 - d + m + \sum_{n=1}^{N} r_{nk})\ln|\boldsymbol{\Lambda}| - \frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\boldsymbol{\Psi}_0^{-1}\boldsymbol{\Lambda}) + \text{const},
\end{aligned}
$$

leading to

$$
\begin{aligned}
\nu_k &= \nu_0 + 1 + m + \sum_{n=1}^{N} r_{nk}, \\
\boldsymbol{\Psi}_k^{-1} &= \boldsymbol{\Psi}_0^{-1} + \sum_{n=1}^{N} r_{nk}\big[(\bar{\boldsymbol{h}}_n - \bar{\mathbf{W}}_k\boldsymbol{x}_n - \bar{\boldsymbol{b}}_k)(\bar{\boldsymbol{h}}_n - \bar{\mathbf{W}}_k\boldsymbol{x}_n - \bar{\boldsymbol{b}}_k)^{\top} + \boldsymbol{\Sigma}_h^{-1}\big] \\
&\quad + (\bar{\mathbf{W}}_k - \mathbf{V}_0)\mathbf{P}_0(\bar{\mathbf{W}}_k - \mathbf{V}_0)^{\top} + \kappa_0(\bar{\boldsymbol{b}} - \boldsymbol{m}_{b_0})(\bar{\boldsymbol{b}} - \boldsymbol{m}_{b_0})^{\top}.
\end{aligned}
$$

4.3.5 Updating the projection matrix A

For the projection matrix $\mathbf{A}$ the parameter updates are derived by

$$
\begin{aligned}
\ln q(\mathbf{A}|\boldsymbol{\Delta}) &= \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{c},\boldsymbol{h})}\left[\ln p(\mathbf{Y}|\mathbf{Z},\mathbf{H},\mathbf{A},\boldsymbol{c},\boldsymbol{\Delta}) + \ln p(\mathbf{A}|\boldsymbol{\Delta})\right] \\
&= \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{c},\boldsymbol{h})}\left[\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\ln\mathcal{N}(\boldsymbol{y}_n|\mathbf{A}_k\boldsymbol{h}_n+\boldsymbol{c}_k,\boldsymbol{\Delta})\right] + \sum_{k=1}^{K}\ln\mathcal{MN}(\mathbf{A}_k|\mathbf{M}_0,\mathbf{K}_0,\boldsymbol{\Delta}) \\
&= -\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}(\boldsymbol{y}_n-\mathbf{A}_k\bar{\boldsymbol{h}}_n-\bar{\boldsymbol{c}}_k)^\top\boldsymbol{\Delta}(\boldsymbol{y}_n-\mathbf{A}_k\bar{\boldsymbol{h}}_n-\bar{\boldsymbol{c}}_k) + r_{nk}\,\mathrm{tr}(\mathbf{A}^\top\boldsymbol{\Delta}\mathbf{A}\boldsymbol{\Sigma}_h^{-1}) \\
&\quad -\frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}(\mathbf{K}_0(\mathbf{A}_k-\mathbf{M}_0)^\top\boldsymbol{\Delta}(\mathbf{A}_k-\mathbf{M}_0)) + \text{const} \\
&= -\frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}(\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\mathbf{Y} - 2\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\mathbf{A}_k\bar{\mathbf{H}} - 2\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\bar{\mathbf{C}}_k \\
&\qquad\qquad + \mathbf{R}\bar{\mathbf{H}}^\top\mathbf{A}_k^\top\boldsymbol{\Delta}\mathbf{A}_k\bar{\mathbf{H}} + 2\mathbf{R}\bar{\mathbf{H}}^\top\mathbf{A}_k^\top\boldsymbol{\Delta}\bar{\mathbf{C}}_k + \mathbf{R}\bar{\mathbf{C}}_k^\top\boldsymbol{\Delta}\bar{\mathbf{C}}_k + \mathbf{R}\mathbf{A}^\top\boldsymbol{\Delta}\mathbf{A}\sum_{n=1}^{N}\boldsymbol{\Sigma}_h^{-1}) \\
&\quad -\frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}(\mathbf{K}_0\mathbf{A}_k^\top\boldsymbol{\Delta}\mathbf{A}_k - 2\mathbf{K}_0\mathbf{A}^\top\boldsymbol{\Delta}\mathbf{M}_0 + \mathbf{K}_0\mathbf{M}_0^\top\boldsymbol{\Delta}\mathbf{M}_0) + \text{const} \\
&= -\frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}(-2\mathbf{A}_k^\top\boldsymbol{\Delta}(\mathbf{M}_0\mathbf{K}_0 + \mathbf{Y}\mathbf{R}\bar{\mathbf{H}}^\top - \bar{\mathbf{C}}_k\mathbf{R}\bar{\mathbf{H}}^\top) \\
&\qquad\qquad + \mathbf{A}_k^\top\boldsymbol{\Delta}\mathbf{A}_k(\mathbf{K}_0 + \mathbf{H}\mathbf{R}\mathbf{H}^\top + \mathbf{R}\sum_{n=1}^{N}\boldsymbol{\Sigma}_h^{-1})) + \text{const},
\end{aligned}
$$

resulting in

$$
\begin{aligned}
\mathbf{K} &= \mathbf{K}_0 + \mathbf{H}\mathbf{R}\mathbf{H}^\top + \mathbf{R}\sum_{n=1}^{N}\boldsymbol{\Sigma}_h^{-1}, \\
\mathbf{M} &= (\mathbf{M}_0\mathbf{K}_0 + \mathbf{Y}\mathbf{R}\mathbf{H}^\top - \bar{\mathbf{C}}_k\mathbf{R}\mathbf{H}^\top)(\mathbf{K}_0 + \mathbf{H}\mathbf{R}\mathbf{H}^\top + \mathbf{R}\sum_{n=1}^{N}\boldsymbol{\Sigma}_h^{-1})^{-1}.
\end{aligned}
$$

4.3.6 Updating the translation vector c

To update g and $\mathbf{m}_c$ we derive

$$\begin{aligned}
\ln q(\boldsymbol{c}|\boldsymbol{\Delta}) &= \mathbb{E}_{q(\boldsymbol{z},\mathbf{A},\boldsymbol{h})}\left[\ln p(\mathbf{Y}|\mathbf{Z},\mathbf{H},\mathbf{A},\boldsymbol{c},\boldsymbol{\Delta})\right] + \ln p(\boldsymbol{c}|\boldsymbol{\Delta}) \\
&= -\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}(\boldsymbol{y}_n - \bar{\mathbf{A}}_k\bar{\boldsymbol{h}}_n - \boldsymbol{c}_k)^\top \boldsymbol{\Delta}(\boldsymbol{y}_n - \bar{\mathbf{A}}_k\bar{\boldsymbol{h}}_n - \boldsymbol{c}_k) \\
&\quad -\frac{1}{2}\sum_{k=1}^{K}(\boldsymbol{c}_k - \boldsymbol{m}_{c_0})^\top g_0\boldsymbol{\Delta}(\boldsymbol{c}_k - \boldsymbol{m}_{c_0}) + \text{const} \\
&= -\frac{1}{2}\left[\sum_{k=1}^{K}(g_0 + \sum_{n=1}^{N} r_{nk})\boldsymbol{c}_k^\top\boldsymbol{\Delta}\boldsymbol{c}_k - 2\sum_{k=1}^{K}\boldsymbol{c}_k^\top\boldsymbol{\Delta}(g_0\boldsymbol{m}_{c_0} + \sum_{n=1}^{N} r_{nk}(\boldsymbol{y}_n - \bar{\mathbf{A}}_k\bar{\boldsymbol{h}}_n))\right] + \text{const},
\end{aligned}$$

resulting in

$$\begin{aligned}
g &= g_0 + \sum_{n=1}^{N} r_{nk}, \\
\boldsymbol{m}_c &= \frac{g_0\boldsymbol{m}_{c_0} + \sum_{n=1}^{N} r_{nk}(\boldsymbol{y}_n - \bar{\mathbf{A}}_k\bar{\boldsymbol{h}}_n)}{g_0 + \sum_{n=1}^{N} r_{nk}}.
\end{aligned}$$

4.3.7 Updating the latent variable h

For $q(\mathbf{h}_n)$ we derive very similar to the previous parameters

$$\begin{aligned}
\ln q(\mathbf{h}_n) &= \mathbb{E}_{\boldsymbol{\Theta}}\left[\ln p(\mathbf{y}_n|\mathbf{z}_n,\mathbf{A}\mathbf{h}_n,\boldsymbol{\Delta}) + \ln p(\mathbf{h}_n|\mathbf{z}_n,\mathbf{W}\mathbf{x}_n,\boldsymbol{\Lambda})\right] \\
&= \mathbb{E}\left[\sum^{k} z_{nk}\left[\ln\mathcal{N}(\mathbf{y}_n|\mathbf{z}_n,\mathbf{A}_k\mathbf{h}_n + \mathbf{c}_k,\boldsymbol{\Delta}_k) + \ln\mathcal{N}(\mathbf{h}_n|\mathbf{z}_n,\mathbf{W}_k\mathbf{x}_n + \mathbf{b}_k,\boldsymbol{\Lambda}_k)\right]\right] \\
&= \sum_{k=1}^{K} r_{nk}\big(\mathbf{y}_n^\top\bar{\boldsymbol{\Delta}}_k\mathbf{y}_n - \mathbf{y}_n^\top\bar{\boldsymbol{\Delta}}_k\bar{\mathbf{A}}_k\mathbf{h}_n - \mathbf{y}_n^\top\bar{\boldsymbol{\Delta}}_k\bar{\mathbf{c}}_k - \mathbf{h}_n^\top\bar{\mathbf{A}}_k^\top\bar{\boldsymbol{\Delta}}_k\mathbf{y}_n + \mathbf{h}_n^\top\bar{\mathbf{A}}_k^\top\bar{\boldsymbol{\Delta}}_k\bar{\mathbf{A}}_k\mathbf{h}_n - \bar{\mathbf{c}}_k^\top\bar{\boldsymbol{\Delta}}_k\mathbf{y}_n \\
&\qquad + \mathbf{h}_n^\top\bar{\mathbf{A}}_k^\top\bar{\boldsymbol{\Delta}}_k\bar{\mathbf{c}}_k - \bar{\mathbf{c}}_k^\top\bar{\boldsymbol{\Delta}}_k\bar{\mathbf{A}}_k\mathbf{h}_n + \bar{\mathbf{c}}_k^\top\bar{\boldsymbol{\Delta}}_k\bar{\mathbf{c}}_k + dg^{-1} + \mathbf{K}d + \mathbf{h}_n^\top\bar{\boldsymbol{\Lambda}}\mathbf{h}_n - \mathbf{h}_n^\top\bar{\boldsymbol{\Lambda}}\bar{\mathbf{W}}_k\mathbf{x}_n \\
&\qquad - \mathbf{h}_n^\top\bar{\boldsymbol{\Lambda}}\bar{\mathbf{b}}_k - \mathbf{x}_n^\top\bar{\mathbf{W}}_k^\top\bar{\boldsymbol{\Lambda}}\mathbf{h}_n + \mathbf{x}_n^\top\bar{\mathbf{W}}_k^\top\bar{\boldsymbol{\Lambda}}\bar{\mathbf{W}}_k\mathbf{x}_n \\
&\qquad + \mathbf{x}_n^\top\bar{\mathbf{W}}_k^\top\bar{\boldsymbol{\Lambda}}\bar{\mathbf{b}}_k - \bar{\mathbf{b}}_k^\top\bar{\boldsymbol{\Lambda}}\mathbf{h}_n + \bar{\mathbf{b}}_k^\top\bar{\boldsymbol{\Lambda}}\bar{\mathbf{W}}_k\mathbf{x}_n + \bar{\mathbf{b}}_k^\top\bar{\boldsymbol{\Lambda}}\bar{\mathbf{b}}_k + \kappa^{-1}d + \mathbf{x}_n^\top\mathbf{P}d\mathbf{x}_n\big),
\end{aligned}$$

leading to

$$\begin{aligned}
\boldsymbol{\Sigma}_h &= \left(\bar{\mathbf{A}}_k^\top\bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k + \bar{\boldsymbol{\Lambda}}\right)^{-1}, \\
\mathbf{m}_h &= \boldsymbol{\Sigma}_h\left[(\mathbf{y}^\top - \bar{\mathbf{c}}^\top)\bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k + (\mathbf{x}^\top\bar{\mathbf{W}}^\top + \bar{\mathbf{b}}^\top)\bar{\boldsymbol{\Lambda}}\right].
\end{aligned}$$

4.3.8 Updating the precision matrix $\boldsymbol{\Delta}$

The posterior of $\boldsymbol{\Delta}$ can be updated with

$$
\begin{aligned}
\ln q(\boldsymbol{\Delta}) &= \mathbb{E}_{q(\boldsymbol{z},\mathbf{A},\mathbf{c},\boldsymbol{h})}\left[\ln p(\mathbf{Y}|\mathbf{Z},\mathbf{H},\mathbf{A},\boldsymbol{c},\boldsymbol{\Delta})\right] + \ln p(\boldsymbol{\Delta}) + \mathbb{E}_{q(\mathbf{A},\mathbf{c})}\left[\ln p(\mathbf{A}|\boldsymbol{\Delta}) + \ln p(\boldsymbol{c}|\boldsymbol{\Delta})\right] \\
&= \mathbb{E}_{q(\boldsymbol{z},\mathbf{A},\boldsymbol{c})}\left[\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\ln\mathcal{N}(\boldsymbol{y}_n|\mathbf{A}_k\boldsymbol{h}_n+\boldsymbol{c}_k,\boldsymbol{\Delta})\right] + \sum_{k=1}^{K}\ln\mathcal{W}(\boldsymbol{\Delta}|\boldsymbol{\Theta}_0,\phi_0) \\
&\quad + \mathbb{E}_{q(\mathbf{A})}\left[\sum_{k=1}^{K}\ln\mathcal{MN}(\mathbf{A}_k|\mathbf{M}_0,\mathbf{K}_0,\boldsymbol{\Delta})\right] + \mathbb{E}_{q(\mathbf{c})}\left[\sum_{k=1}^{K}\ln\mathcal{N}(\boldsymbol{c}|\boldsymbol{m}_{c_0},\boldsymbol{g}_0\boldsymbol{\Delta})\right] \\
&= +\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln|\boldsymbol{\Delta}| - \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\mathbb{E}\left[(\boldsymbol{y}_n-\mathbf{A}_k\boldsymbol{h}_n-\boldsymbol{c}_k)^\top\boldsymbol{\Delta}(\boldsymbol{y}_n-\mathbf{A}_k\boldsymbol{h}_n-\boldsymbol{c}_k)\right] \\
&\quad + \frac{m}{2}\sum_{k=1}^{K}\ln|\boldsymbol{\Delta}| - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\mathrm{tr}(\mathbf{K}_0(\mathbf{A}_k-\mathbf{M}_0)^\top\boldsymbol{\Delta}(\mathbf{A}_k-\mathbf{M}_0))\right] \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}\ln|\boldsymbol{\Delta}| - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[(\boldsymbol{c}_k-\boldsymbol{m}_{c_0})^\top(g_0\boldsymbol{\Delta})(\boldsymbol{c}_k-\boldsymbol{m}_{c_0})\right] \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}(\phi_0-d-1)\ln|\boldsymbol{\Delta}| - \frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}(\boldsymbol{\Theta}_0^{-1}\boldsymbol{\Delta}) + \text{const} \\
&= -\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\Big[\,\mathrm{tr}(\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\mathbf{Y} - 2\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\mathbf{A}_k\mathbf{H} - 2\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\mathbf{C}_k \\
&\qquad\qquad + \mathbf{R}\mathbf{H}^\top\mathbf{A}_k^\top\boldsymbol{\Delta}\mathbf{A}_k\mathbf{H} + 2\mathbf{R}\mathbf{H}^\top\mathbf{A}_k^\top\boldsymbol{\Delta}\mathbf{C}_k + \mathbf{R}\mathbf{C}_k^\top\boldsymbol{\Delta}\mathbf{C}_k)\Big] \\
&\quad - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\Big[\,\mathrm{tr}(\mathbf{K}_0\mathbf{A}_k^\top\boldsymbol{\Delta}\mathbf{A}_k - 2\mathbf{K}_0\mathbf{A}^\top\boldsymbol{\Delta}\mathbf{M}_0 + \mathbf{K}_0\mathbf{M}_0^\top\boldsymbol{\Delta}\mathbf{M}_0)\Big] \\
&\quad - \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\Big[(\boldsymbol{c}-\boldsymbol{m}_{c_0})^\top(g_0\boldsymbol{\Delta})(\boldsymbol{c}-\boldsymbol{m}_{c_0})\Big] \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln|\boldsymbol{\Delta}| + \frac{m}{2}\sum_{k=1}^{K}\ln|\boldsymbol{\Delta}| + \frac{1}{2}\sum_{k=1}^{K}\ln|\boldsymbol{\Delta}| \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}(\phi_0-d-1)\ln|\boldsymbol{\Delta}| - \frac{1}{2}\sum_{k=1}^{K}\mathrm{tr}(\boldsymbol{\Theta}_0^{-1}\boldsymbol{\Delta}) + \text{const}
\end{aligned}
$$

$$
\begin{aligned}
= & -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\mathbf{Y} - 2\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\bar{\mathbf{A}}_k\bar{\mathbf{H}} - 2\mathbf{R}\mathbf{Y}^\top\boldsymbol{\Delta}\bar{\mathbf{C}}_k \\
& \qquad + \mathbf{R}\mathbf{H}^\top\bar{\mathbf{A}}_k^\top\boldsymbol{\Delta}\bar{\mathbf{A}}_k\mathbf{H} + 2\mathbf{R}\bar{\mathbf{H}}^\top\bar{\mathbf{A}}_k^\top\boldsymbol{\Delta}\bar{\mathbf{C}}_k + \mathbf{R}\bar{\mathbf{C}}_k^\top\boldsymbol{\Delta}\bar{\mathbf{C}}_k + \mathbf{R}\bar{\mathbf{A}}_k^\top\boldsymbol{\Delta}\bar{\mathbf{A}}_k\sum_{n=1}^{N}\boldsymbol{\Sigma}_h^{-1}) \\
& -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\mathbf{K}_0\bar{\mathbf{A}}_k^\top\boldsymbol{\Delta}\bar{\mathbf{A}}_k - 2\mathbf{K}_0\bar{\mathbf{A}}^\top\boldsymbol{\Delta}\mathbf{M}_0 + \mathbf{K}_0\mathbf{M}_0^\top\boldsymbol{\Delta}\mathbf{M}_0) \\
& -\frac{1}{2}\sum_{k=1}^{K}(\bar{\boldsymbol{c}} - \boldsymbol{m}_{c_0})^\top(g_0\boldsymbol{\Delta})(\bar{\boldsymbol{c}} - \boldsymbol{m}_{c_0}) \\
& +\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln|\boldsymbol{\Delta}| + \frac{m}{2}\sum_{k=1}^{K}\ln|\boldsymbol{\Delta}| + \frac{1}{2}\sum_{k=1}^{K}\ln|\boldsymbol{\Delta}| \\
& +\frac{1}{2}\sum_{k=1}^{K}(\phi_0 - d - 1)\ln|\boldsymbol{\Delta}| - \frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\boldsymbol{\Theta}_0^{-1}\boldsymbol{\Delta}) + \text{const} \\
= & -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\mathbf{R}(\mathbf{Y} - \bar{\mathbf{A}}_k\bar{\mathbf{H}} - \bar{\mathbf{C}}_k)(\mathbf{Y} - \bar{\mathbf{A}}_k\bar{\mathbf{H}} - \bar{\mathbf{C}}_k)^\top\boldsymbol{\Delta}) \\
& -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}((\bar{\mathbf{A}}_k - \mathbf{M}_0)\mathbf{K}_0(\bar{\mathbf{A}}_k - \mathbf{M}_0)^\top\boldsymbol{\Delta}) \\
& -\frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(g_0(\bar{\boldsymbol{c}} - \boldsymbol{m}_{c_0})(\bar{\boldsymbol{c}} - \boldsymbol{m}_{c_0})^\top\boldsymbol{\Delta}) \\
& +\frac{1}{2}\sum_{k=1}^{K}(\phi_0 - d + m + \sum_{n=1}^{N} r_{nk})\ln|\boldsymbol{\Delta}| - \frac{1}{2}\sum_{k=1}^{K}\operatorname{tr}(\boldsymbol{\Theta}_0^{-1}\boldsymbol{\Delta}) + \text{const},
\end{aligned}
$$

yielding the update equations

$$
\begin{aligned}
\phi_k &= \phi_0 + 1 + m + \sum_{n=1}^{N} r_{nk}, \\
\boldsymbol{\Theta}_k^{-1} &= \boldsymbol{\Theta}_0^{-1} + \sum_{n=1}^{N} r_{nk}(\boldsymbol{y}_n - \bar{\mathbf{A}}_k\bar{\boldsymbol{h}}_n - \bar{\boldsymbol{c}}_k)(\boldsymbol{y}_n - \bar{\mathbf{A}}_k\bar{\boldsymbol{h}}_n - \bar{\boldsymbol{c}}_k)^\top + \sum_{n=1}^{N} r_{nk}\bar{\mathbf{A}}_k^\top\boldsymbol{\Delta}\bar{\mathbf{A}}_k\boldsymbol{\Sigma}_h^{-1} \\
&\quad + (\bar{\mathbf{A}}_k - \mathbf{M}_0)\mathbf{K}_0(\bar{\mathbf{A}}_k - \mathbf{M}_0)^\top + g_0(\bar{\boldsymbol{c}} - \boldsymbol{m}_{c_0})(\bar{\boldsymbol{c}} - \boldsymbol{m}_{c_0})^\top.
\end{aligned}
$$

4.3.9 Updating the one-hot latent variable z

Since most terms depend on $\mathbf{z}$, this derivation is more involved, why it is split into parts first, before reassembled in the end.

The derivation is split with

$$
\begin{aligned}
\ln q(\mathbf{z}) &= \mathbb{E}_{h,A,C,\Delta,W,b,\Lambda,\pi}\Big[\ln\left[p(\mathbf{y}|\mathbf{h},\mathbf{z},\mathbf{A},\mathbf{c},\boldsymbol{\Delta})p(\mathbf{h}|\mathbf{x},\mathbf{z},\mathbf{W},\mathbf{b},\boldsymbol{\Lambda})p(\mathbf{x}|\mathbf{z},\boldsymbol{\eta},\boldsymbol{\Gamma})p(\mathbf{z}|\boldsymbol{\pi})\right]\Big] \\
&= \mathbb{E}_{h,A,C,\Delta,W,b,\Lambda,\pi}\Big[\underbrace{+\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\ln\mathcal{N}(\mathbf{y}_n|\mathbf{A}_k\mathbf{h}_n+\mathbf{c}_k,\boldsymbol{\Delta})}_{(1)} \\
&\qquad \underbrace{+\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\ln\mathcal{N}(\mathbf{h}_n|\mathbf{W}_k\mathbf{x}_n+\mathbf{b}_k,\boldsymbol{\Lambda})}_{(2)} \\
&\qquad +\sum_{k=1}^{K}\underbrace{\sum_{n=1}^{N} z_{nk}\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\eta}_k,\boldsymbol{\Gamma}_k)}_{(3)} \\
&\qquad +\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\ln\pi_k\Big].
\end{aligned}
$$

Part (1) can be written as

$$
\begin{aligned}
&\mathbb{E}_{A,h,c,\Delta}\Big[\mathbf{y}_n^\top \boldsymbol{\Delta}\mathbf{y}_n - \mathbf{y}_n^\top \boldsymbol{\Delta}\mathbf{A}_k\mathbf{h}_n - \mathbf{y}_n^\top \boldsymbol{\Delta}\mathbf{c}_k - \mathbf{h}_n^\top \mathbf{A}_k^\top \boldsymbol{\Delta}\mathbf{y}_n \\
&\qquad + \mathbf{h}_n^\top \mathbf{A}_k^\top \boldsymbol{\Delta}\mathbf{A}_k\mathbf{h}_n + \mathbf{h}_n^\top \mathbf{A}_k^\top \boldsymbol{\Delta}\mathbf{c}_k - \mathbf{c}_k^\top \boldsymbol{\Delta}\mathbf{y}_n + \mathbf{c}_k^\top \boldsymbol{\Delta}\mathbf{A}_k\mathbf{h}_n + \mathbf{c}_k^\top \boldsymbol{\Delta}\mathbf{c}_k\Big] \\
&= \mathbf{y}_n^\top \bar{\boldsymbol{\Delta}}\mathbf{y}_n - \mathbf{y}_n^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\bar{\mathbf{h}}_n - \mathbf{y}_n^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{c}}_k - \bar{\mathbf{h}}_n^\top \bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\mathbf{y}_n \\
&\quad + \mathbb{E}_{h,A,\Delta}[\mathbf{h}_n^\top \mathbf{A}_k^\top \boldsymbol{\Delta}\mathbf{A}_k\mathbf{h}_n] + \bar{\mathbf{h}}_n^\top \bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{c}}_k - \bar{\mathbf{c}}_k^\top \bar{\boldsymbol{\Delta}}\mathbf{y}_n + \bar{\mathbf{c}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\bar{\mathbf{h}}_n + \mathbb{E}_{c,\Delta}[\mathbf{c}_k^\top \boldsymbol{\Delta}\mathbf{c}_k] \\
&= \mathbf{y}_n^\top \bar{\boldsymbol{\Delta}}\mathbf{y}_n - \mathbf{y}_n^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\bar{\mathbf{h}}_n - \mathbf{y}_n^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{c}}_k - \bar{\mathbf{h}}_n^\top \bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\mathbf{y}_n \\
&\quad + \bar{\mathbf{h}}_n^\top \mathbf{K}d\bar{\mathbf{h}}_n + \mathrm{tr}(\mathbf{K}d\boldsymbol{\Sigma}_h^{-1}) + \bar{\mathbf{h}}_n^\top \bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\bar{\mathbf{h}}_n + \mathrm{tr}(\bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\boldsymbol{\Sigma}_h^{-1}) + \bar{\mathbf{h}}_n^\top \bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{c}}_k \\
&\quad - \bar{\mathbf{c}}_k^\top \bar{\boldsymbol{\Delta}}\mathbf{y}_n + \bar{\mathbf{c}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\bar{\mathbf{h}}_n + \bar{\mathbf{c}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{c}}_k + \mathrm{tr}(g^{-1}\mathbf{I}) \\
&= (\mathbf{y}_n - \bar{\mathbf{A}}_k\bar{\mathbf{h}}_n - \bar{\mathbf{c}}_k)^\top \bar{\boldsymbol{\Delta}}(\mathbf{y}_n - \bar{\mathbf{A}}_k\bar{\mathbf{h}}_n - \bar{\mathbf{c}}_k) + \bar{\mathbf{h}}_n^\top \mathbf{K}d\bar{\mathbf{h}}_n + \mathrm{tr}(\mathbf{K}d\boldsymbol{\Sigma}_h^{-1}) + g^{-1}d + \mathrm{tr}(\bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\boldsymbol{\Sigma}_h^{-1}).
\end{aligned}
$$

With $\boldsymbol{\Delta}$ being a symmetric precision matrix we can define

$$
\begin{aligned}
\mathbb{E}_A[\mathbf{A}_k^\top \boldsymbol{\Delta}\mathbf{A}_k] &= \mathbf{K}\,\mathrm{tr}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}^\top) + \bar{\mathbf{A}}_k^\top \boldsymbol{\Delta}\bar{\mathbf{A}}_k \\
&= \mathbf{K}\,\mathrm{tr}(\mathbf{I}_d) + \bar{\mathbf{A}}_k^\top \boldsymbol{\Delta}\bar{\mathbf{A}}_k \\
&= \mathbf{K}d + \bar{\mathbf{A}}_k^\top \boldsymbol{\Delta}\bar{\mathbf{A}}_k,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{h,A,\Delta}[\mathbf{h}_n^\top \mathbf{A}_k^\top \boldsymbol{\Delta}\mathbf{A}_k\mathbf{h}_n] &= \mathbb{E}_{h,\Delta}[\mathbf{h}_n^\top \mathbb{E}_A[\mathbf{A}_k^\top \boldsymbol{\Delta}\mathbf{A}_k]\mathbf{h}_n] \\
&= \mathbb{E}_{h,\Delta}[\mathbf{h}_n^\top (\mathbf{K}d + \bar{\mathbf{A}}_k^\top \boldsymbol{\Delta}\bar{\mathbf{A}}_k)\mathbf{h}_n] \\
&= \mathbb{E}_{h,\Delta}[\mathbf{h}_n^\top \mathbf{K}d\mathbf{h}_n] + \mathbb{E}_{h,\Delta}[\mathbf{h}_n^\top \bar{\mathbf{A}}_k^\top \boldsymbol{\Delta}\bar{\mathbf{A}}_k\mathbf{h}_n] \\
&= \bar{\mathbf{h}}_n^\top \mathbf{K}d\bar{\mathbf{h}}_n + \mathrm{tr}(\mathbf{K}d\boldsymbol{\Sigma}_h^{-1}) + \bar{\mathbf{h}}_n^\top \bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\bar{\mathbf{h}}_n + \mathrm{tr}(\bar{\mathbf{A}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{A}}_k\boldsymbol{\Sigma}_h^{-1}),
\end{aligned}
$$

and

$$
\mathbb{E}_{c,\Delta}[\mathbf{c}_k^\top \boldsymbol{\Delta}\mathbf{c}_k] = \bar{\mathbf{c}}_k^\top \bar{\boldsymbol{\Delta}}\bar{\mathbf{c}}_k + \mathrm{tr}(g^{-1}\mathbf{I}).
$$

Part (2) can be written as

$$
\begin{aligned}
&\mathbb{E}_{h,W,b,\Lambda}\left[-\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}(\mathbf{h}_n-\mathbf{W}_k\mathbf{x}_n-\mathbf{b}_k)^\top\mathbf{\Lambda}(\mathbf{h}_n-\mathbf{W}_k\mathbf{x}_n-\mathbf{b}_k)\right]\\
&=\mathbb{E}_{h,W,b,\Lambda}\Bigg[-\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}(\mathbf{h}_n^\top\mathbf{\Lambda}\mathbf{h}_n-\mathbf{h}_n^\top\mathbf{\Lambda}\mathbf{W}_k\mathbf{x}_n-\mathbf{h}_n^\top\mathbf{\Lambda}\mathbf{b}_k-\mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{\Lambda}\mathbf{h}_n\\
&\qquad\qquad+\mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{\Lambda}\mathbf{W}_k\mathbf{x}_n+\mathbf{x}_n^\top\mathbf{W}_k^\top\mathbf{\Lambda}\mathbf{b}_k-\mathbf{b}_k^\top\mathbf{\Lambda}\mathbf{h}_n+\mathbf{b}_k\mathbf{\Lambda}\mathbf{W}_k\mathbf{x}_n+\mathbf{b}_k^\top\mathbf{\Lambda}\mathbf{b}_k)\Bigg]\\
&=-\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}(\,\mathbb{E}_{h,\Lambda}\left[\mathbf{h}_n^\top\mathbf{\Lambda}\mathbf{h}_n\right]-\bar{\mathbf{h}}_n^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{W}}_k\bar{\mathbf{x}}_n-\bar{\mathbf{h}}_n^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{b}}_k-\mathbf{x}_n^\top\bar{\mathbf{W}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{h}}_n\\
&\qquad\qquad+\mathbf{x}_n^\top\,\mathbb{E}_{W,\Lambda}\left[\mathbf{W}_k^\top\mathbf{\Lambda}\mathbf{W}_k\right]\mathbf{x}_n+\mathbf{x}_n^\top\bar{\mathbf{W}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{b}}_k-\bar{\mathbf{b}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{h}}_n+\bar{\mathbf{b}}_k\bar{\mathbf{\Lambda}}\bar{\mathbf{W}}_k\mathbf{x}_n+\mathbb{E}_{b,\Lambda}\left[\mathbf{b}_k^\top\mathbf{\Lambda}\mathbf{b}_k\right])\\
&=-\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}(\bar{\mathbf{h}}_n^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{h}}_n+\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Sigma}_h^{-1})-\bar{\mathbf{h}}_n^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{W}}_k\bar{\mathbf{x}}_n-\bar{\mathbf{h}}_n^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{b}}_k-\mathbf{x}_n^\top\bar{\mathbf{W}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{h}}_n\\
&\qquad\qquad+\mathbf{x}_n^\top(\mathbf{P}d+\bar{\mathbf{W}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{W}}_k)\mathbf{x}_n+\mathbf{x}_n^\top\bar{\mathbf{W}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{b}}_k-\bar{\mathbf{b}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{h}}_n+\bar{\mathbf{b}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{W}}_k\mathbf{x}_n+\bar{\mathbf{b}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{b}}_k+\kappa^{-1}d)\\
&=-\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}((\bar{\mathbf{h}}_n-\bar{\mathbf{W}}_k\mathbf{x}_n-\bar{\mathbf{b}}_k)^\top\bar{\mathbf{\Lambda}}(\bar{\mathbf{h}}_n-\bar{\mathbf{W}}_k\mathbf{x}_n-\bar{\mathbf{b}}_k)+\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Sigma}_h^{-1})+\kappa^{-1}d+d\mathbf{x}_n^\top\mathbf{P}\mathbf{x}_n)\\
&=-\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}((\bar{\mathbf{h}}_n-\bar{\mathbf{W}}_k\mathbf{x}_n-\bar{\mathbf{b}}_k)^\top\bar{\mathbf{\Lambda}}(\bar{\mathbf{h}}_n-\bar{\mathbf{W}}_k\mathbf{x}_n-\bar{\mathbf{b}}_k)+\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Sigma}_h^{-1})+\kappa^{-1}d+d\mathbf{x}_n^\top\mathbf{P}\mathbf{x}_n),
\end{aligned}
$$

where we used the identities

$$\mathbb{E}_{h,\Lambda}[\mathbf{h}_n^\top\mathbf{\Lambda}\mathbf{h}_n]=\bar{\mathbf{h}}_n^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{h}}_n+\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Sigma}_h^{-1}),$$

$$
\begin{aligned}
\mathbb{E}_{W,\Lambda}[\mathbf{W}_k^\top\mathbf{\Lambda}\mathbf{W}_k]&=\mathbf{P}\,\mathrm{tr}(\mathbf{\Lambda}^{-\top}\mathbf{\Lambda})+\bar{\mathbf{W}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{W}}_k\\
&=\mathbf{P}d+\bar{\mathbf{W}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{W}}_k,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}_{b,\Lambda}[\mathbf{b}_k^\top\mathbf{\Lambda}\mathbf{b}_k]&=\bar{\mathbf{b}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{b}}_k+\mathrm{tr}(\mathbf{\Lambda}\kappa^{-1}\mathbf{\Lambda}^{-1})\\
&=\bar{\mathbf{b}}_k^\top\bar{\mathbf{\Lambda}}\bar{\mathbf{b}}_k+\kappa^{-1}d.
\end{aligned}
$$

Part (3) can be written as

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\eta},\boldsymbol{\Gamma}}\left[\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\ln p(\mathbf{x}_n|\boldsymbol{\eta}_k,\boldsymbol{\Gamma}_k)\right]\\
&=\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\left[\frac{1}{2}\,\mathbb{E}[\ln|\boldsymbol{\Gamma}_k|]-\frac{D}{2}\ln(2\pi)-\frac{1}{2}\,\mathbb{E}_{\boldsymbol{\eta}_k,\boldsymbol{\Gamma}_k}[(\mathbf{x}_n-\boldsymbol{\eta}_k)^\top\boldsymbol{\Gamma}_k(\mathbf{x}_n-\boldsymbol{\eta}_k)]\right]\\
&=\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\left[\frac{1}{2}\,\mathbb{E}[\ln|\boldsymbol{\Gamma}_k|]-\frac{D}{2}\ln(2\pi)-\frac{1}{2}\,\mathbb{E}_{\boldsymbol{\eta}_k,\boldsymbol{\Gamma}_k}[\mathbf{x}_n^\top\boldsymbol{\Gamma}_k\mathbf{x}_n-\mathbf{x}_n^\top\boldsymbol{\Gamma}_k\boldsymbol{\eta}_k-\boldsymbol{\eta}_k^\top\boldsymbol{\Gamma}_k\mathbf{x}_n+\boldsymbol{\eta}_k^\top\boldsymbol{\Gamma}_k\boldsymbol{\eta}_k]\right]\\
&=\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\left[\frac{1}{2}\,\mathbb{E}[\ln|\boldsymbol{\Gamma}_k|]-\frac{D}{2}\ln(2\pi)-\frac{1}{2}[(\mathbf{x}_n-\bar{\boldsymbol{\eta}}_k)^\top\bar{\boldsymbol{\Gamma}}_k(\mathbf{x}_n-\bar{\boldsymbol{\eta}}_k)]\right],
\end{aligned}
$$

with

$$
\begin{aligned}
\ln\tilde{\boldsymbol{\Lambda}}_k &\equiv \mathbb{E}[\ln\det\boldsymbol{\Lambda}_k]\\
&=\sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right)+D\ln 2+\ln\det\mathbf{W}_k,\\
\ln\tilde{\pi}_k &\equiv \mathbb{E}[\ln\pi_k]=\psi(u_k)-\psi(\hat{u}).
\end{aligned}
$$

The results from (1), (2), and (3) can now be combined to

$$
\begin{aligned}
\ln q(\mathbf{z}) &= \mathbb{E}_{h,A,C,\Delta,W,b,\Lambda,\pi}\left[\ln\left[p(\mathbf{y}|\mathbf{h},\mathbf{z},\mathbf{A},\mathbf{c},\boldsymbol{\Delta})p(\mathbf{h}|\mathbf{x},\mathbf{z},\mathbf{W},\mathbf{b},\boldsymbol{\Lambda})p(\mathbf{z}|\boldsymbol{\pi})\right]\right]\\
&= (1)+(2)+(3)+\mathbb{E}[\ln\pi_k].
\end{aligned}
$$

4.3.10 Updating η and Γ

For the Normal-Wishart prior over $\boldsymbol{\eta}$ and $\boldsymbol{\Gamma}$, we can see from the Variational Mixture of Gaussians example in [6] that

$$q(\boldsymbol{\eta}_k, \boldsymbol{\Gamma}_k) = \mathcal{N}(\boldsymbol{\eta}_k|\boldsymbol{\gamma}_k, (\beta\boldsymbol{\Gamma}_k)^{-1})\mathcal{W}(\boldsymbol{\Gamma}_k|\boldsymbol{\Xi}_k, \boldsymbol{\xi}_k),$$

with

$$\begin{aligned}
\beta_k &= \beta_0 + N_k, \\
\boldsymbol{\gamma}_k &= \frac{1}{\beta_k}(\beta_0\boldsymbol{\gamma}_0 + N_k\bar{\mathbf{x}}_k), \\
\boldsymbol{\Xi}_k^{-1} &= \boldsymbol{\Xi}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\mathbf{x}}_k - \boldsymbol{\gamma}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\gamma}_0)^\top, \\
\boldsymbol{\xi}_k &= \boldsymbol{\xi}_0 + N_k + 1,
\end{aligned}$$

and

$$\begin{aligned}
N_k &= \sum_{n=1}^{N} r_{nk}, \\
\bar{\mathbf{x}}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n, \\
\mathbf{S}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\top.
\end{aligned}$$

We now have the complete set of parameter update equations, to learn our model on a given dataset.

4.4 Validation Examples

4.4.1 Cosmic microwave background dataset

To evaluate our algorithm, we apply it to the cosmic microwave background dataset [4], which has an input and output dimensionality of 1, in Figure 4.3. We use 50 mixture components, and a hidden space with dimensionality 3.

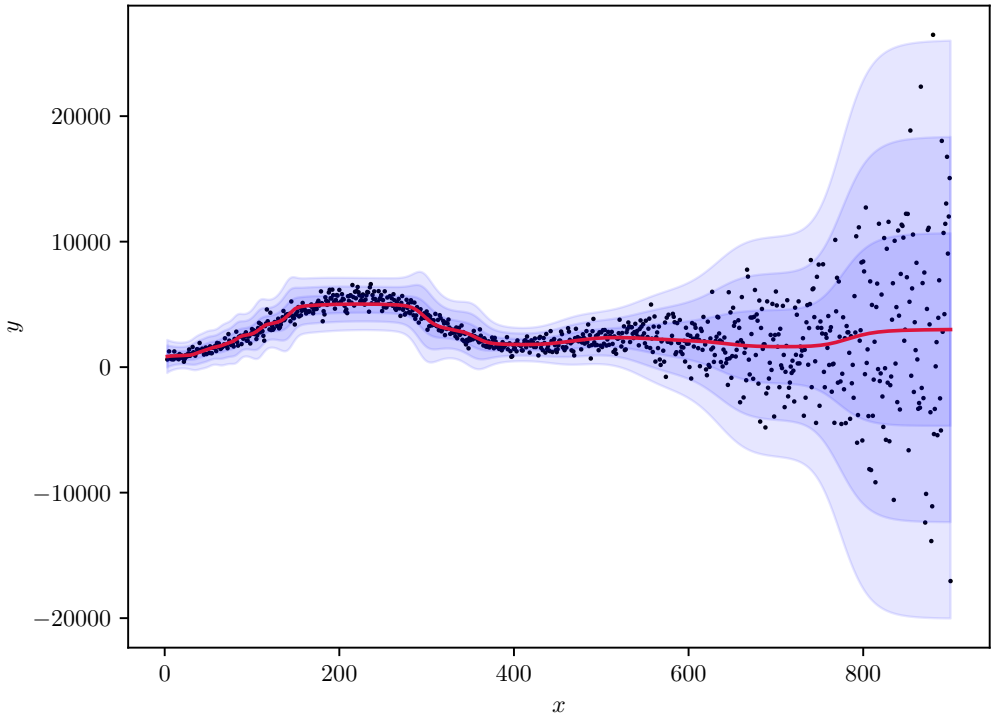

Figure 4.3: The cosmic microwave background dataset approximated by the proposed algorithm. The prediction mean is given by the red curve, while the 1,2,3 standard deviation is shown in purple.

4.4.2 Regression on a Non-Linear Toy Dataset

The next example is a dataset, created by the function

$$y = \sin(2x) + 2\exp(-16x^2) + N(0, 0.16).$$

As shown in 4.4,we are able to successfully model the dataset with 25 mixture components.

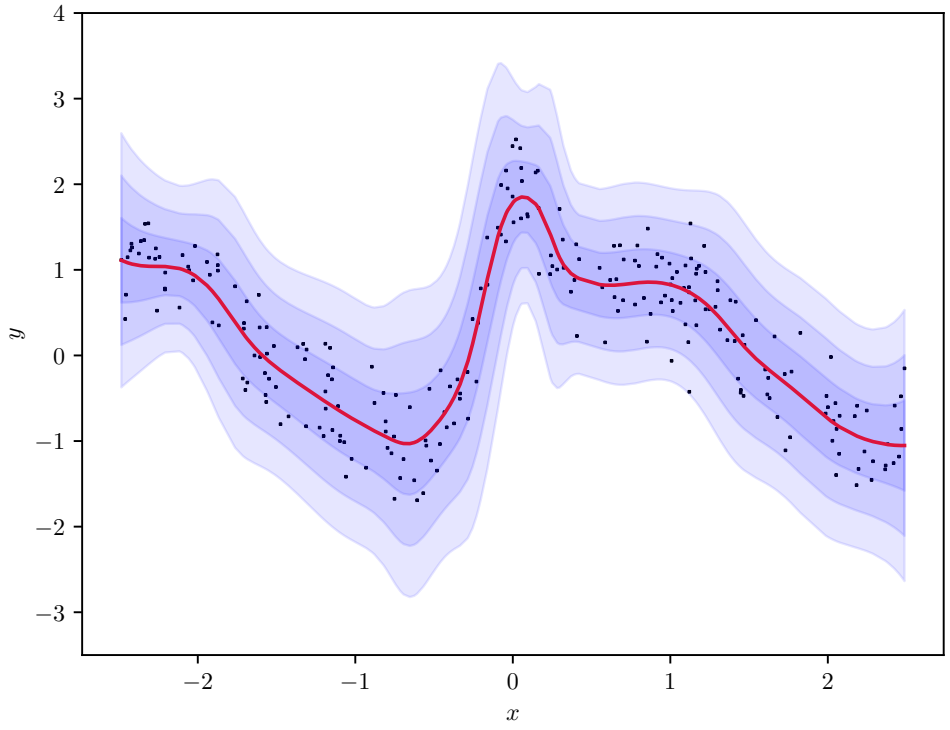

Figure 4.4: The derived algorithm applied to a non-linear toy dataset, with 25 mixture components. The prediction mean is indicated in red, and the 1,2,3 standard deviation is indicated in purple.

5 Conclusion

We now conclude this thesis by giving a summary of the presented work, and offer an outlook for potential future research.

5.1 Summary

We have shown how the PCA method is able to effectively reduce the dimensionality of a given dataset, by projecting the datapoints onto a principal subspace with a given dimension, such that the loss of information is minimal. We then introduced the probabilistic formulation of PCA, called PPCA. This probabilistic treatment allows extending the method to use the fully Bayesian framework and incorporating mixture models. In chapter 2.3 we showed that BPCA can automatically infer the dimensionality of the principal subspace from the data, and that an EM algorithm can be derived to efficiently compute the solution, even for big datasets. The MPPCA algorithm showed how to incorporate mixture models for PPCA, allowing to operate on complex datasets. We then addressed how the mixture model approach can also be used with BPCA. BMPCA was able to use a mixture of Bayesian principal component analyzers, while automatically inferring the dimensionality of the latent space, as well as the number of models.

We presented the idea of Variational Locally Projected Regression, showing how it is possible to make use of hidden structure, by projecting a given dataset onto a higher dimensional latent space. Followed by a derivation of an algorithm with VI, that can compute solutions even for complex datasets. The derived algorithm has then been sucessfully used, to model two example datasets.

5.2 Future Work

In future research, several topics can still be addressed.

The method introduced in this thesis, makes use of mixture models where the number of mixtures is a hyperparameter. To eliminate the need for this hyperparameter, an extension to infinite mixture models could be formulated.

Furthermore, similar to Deep Neural Networks (DNN), instead of using only a single projection layer, this method could be extended to make use of multiple hidden layers.

To improve the efficiency of the training phase, the stochastic updates could be implemented over batches, such that scaling to larger datasets becomes feasible.

Due to using a Bayesian treatment, the prior distributions require several hyperparameters, which have to be defined and rely on educated manual tuning. A method for optimizing these parameters could increase the predictive performance and make the method easier to use.

Finally the method could be applied to an inverse dynamics learning task, and tested on a real platform.

Bibliography

[1] H. Abdulsamad, P. Nickl, P. Klink, and J. Peters. A variational infinite mixture for probabilistic inverse dynamics learning. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pages 4216–4222. IEEE, 2021.

[2] H. Attias. Inferring parameters and structure of latent variable models by variational bayes. arXiv preprint arXiv:1301.6676, 2013.

[3] D. Barber. Bayesian reasoning and machine learning. Cambridge University Press, 2012.

[4] C. L. Bennett, M. Halpern, G. Hinshaw, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, L. Page, D. N. Spergel, G. S. Tucker, E. Wollack, E. L. Wright, C. Barnes, M. R. Greason, R. S. Hill, E. Komatsu, M. R. Nolta, N. Odegard, H. V. Peiris, L. Verde, and J. L. Weiland. First-year wilkinson microwave anisotropy probe (WMAP) observations: Preliminary maps and basic results. The Astrophysical Journal Supplement Series, 148(1):1–27, sep 2003.

[5] C. M. Bishop. Bayesian pca. Advances in neural information processing systems, pages 382–388, 1999.

[6] C. M. Bishop. Pattern recognition and machine learning. springer, 2006.

[7] C. M. Bishop et al. Neural Networks for Pattern Recognition. Oxford university press, 1995.

[8] C. M. Bishop and J. M. Winn. Non-linear bayesian image modelling. In European Conference on Computer Vision, pages 3–17. Springer, 2000.

[9] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635, 2018.

[10] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. Bayesian Data Analysis. 2013.

[11] I. Goodfellow, Y. Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[12] H. Hotelling. Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6):417, 1933.

[13] I. T. Jolliffe. Principal Component Analysis. New York: Springer-Verlag, 1986.

[14] David J. C. MacKay. Bayesian neural networks and density networks. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 354(1):73–80, 1995.

[15] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation, volume 24, pages 109–165. Elsevier, 1989.

[16] K. P. Murphy. Machine learning: a probabilistic perspective. MIT press, 2012.

[17] P. Nickl. Bayesian Inference for Regression Models using Nonparametric Infinite Mixtures. PhD thesis, 2019.

[18] Sam Roweis. Em algorithms for pca and spca. Advances in neural information processing systems, 10:626–632, 1997.

[19] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. Neural computation, 11(2):443–482, 1999.

[20] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3):611–622, 1999.

[21] C. K. Williams and C. E. Rasmussen. Gaussian processes for machine learning, volume 2. MIT press Cambridge, MA, 2006.

List of Figures

Acronyms

ARD Automatic Relevance Detection

BMPCA Bayesian Mixture Principal Component Analysis

BPCA Bayesian Principal Component Analysis

DNN Deep Neural Networks

DP-GLM Dirichlet Process Mixtures of Generalized Linear Models

EM Expectation Maximization

GEM Generalized Expectation Maximization

MPPCA Mixture of Probabilistic Principal Component Analysers

PCA Principal Component Analysis

PPCA Probablistic Principal Component Analysis

VI Variational Inference